Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Systems Biology

FALCON: A Toolbox for the Fast Contextualisation of Logical Networks

Sébastien De Landtsheer [1,†], Panuwat Trairatphisan [1,†], Philippe Lucarelli [1] and Thomas Sauter [1,*]

[1]Systems Biology Group, Life Sciences Research Unit, University of Luxembourg, Belvaux, Luxembourg

*To whom correspondence should be addressed.

[†]These authors contributed equally to the manuscript.

Associate Editor: Prof. Bonnie Berger

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Mathematical modelling of regulatory networks allows for the discovery of knowledge at the system level. However, existing modelling tools are often computation-heavy and do not offer intuitive ways to explore the model, to test hypotheses or to interpret the results biologically.

Results: We have developed a computational approach to contextualise logical models of regulatory networks with biological measurements based on a probabilistic description of rule-based interactions between the different molecules. Here, we propose a Matlab toolbox, FALCON, to automatically and efficiently build and contextualise networks, which includes a pipeline for conducting parameter analysis, knockouts, and easy and fast model investigation. The contextualised models could then provide qualitative and quantitative information about the network and suggest hypotheses about biological processes.

Availability and implementation: FALCON is freely available for non-commercial users on GitHub under the GPLv3 licence. The toolbox, installation instructions, full documentation and test datasets are available at https://github.com/sysbiolux/FALCON. FALCON runs under Matlab (MathWorks) and requires the Optimization Toolbox.

Contact: thomas.sauter@uni.lu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The functional characteristics of eukaryotic cells are largely determined by the properties of their regulatory networks. Notwithstanding the vast amount of biological data accumulated over the past decades, a global model of the way these networks determine the phenotypes of both healthy and diseased cells remains elusive. One goal of systems biology is to understand these networks at the highest possible functional level, for example to devise therapeutic strategies for treating patients affected by diseases like cancer.

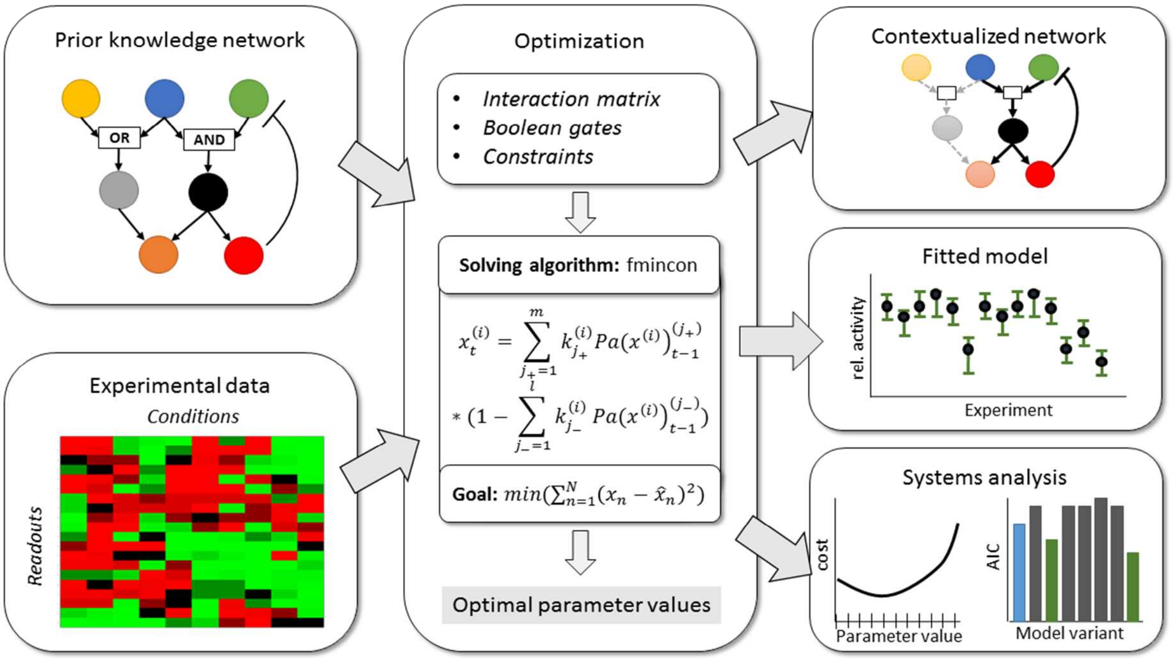

Figure 1. The FALCON pipeline. Prior knowledge network and experimental data are combined to generate a network optimization problem. After the optimization process, the properties of the optimal network are then analyzed.

Numerous mathematical approaches exist to optimize and train regulatory network models against steady-state experimental data (Villaverde and Banga, 2013). Of these, logical models (Le Novère, 2015) are of particular interest, as they are able to capture essential features of the system being modelled and generate biological insights, while requiring less prior knowledge and experimental observations than differential equation models (Morris *et al.*, 2010). Some successful applications include the logical models of yeast cell-cycle protein network (Li *et al.*, 2004), gene regulatory networks (Mendoza *et al.*, 1998), signalling networks (Saez-Rodriguez *et al.*, 2007). In addition, logical models are in general more powerful than statistical models, as they incorporate the relational information embedded in the network structure, while statistical models aiming at reverse-engineering biological networks from high-throughput data implicitly consider all possible topologies (Bansal *et al.*, 2007).

In logical models of systems at steady-state, nodes represent the degree of activation of the constituents of the system at equilibrium and edges represent the logical functions between nodes. These functions can be either linear or non-linear functions of the parent nodes and are combinations of the fundamental '*AND*', '*OR*', and '*NOT*' Boolean functions.

While Binary Boolean models (Kauffman, 1969) only consider full activation or complete absence, more quantitative approaches, for instance, Probabilistic Boolean Networks (PBNs) (Trairatphisan *et al.*, 2013) and Dynamic Bayesian Networks (DBNs) (Lähdesmäki *et al.*, 2006) can account for intermediate or continuous activation values and allow the integration of data uncertainty. These approaches are usually analysed by Monte Carlo approaches (Trairatphisan *et al.*, 2014; Mizera *et al.*, 2016), which can be computationally demanding or non-intuitive to use. Here, we propose a tool called FALCON to efficiently contextualize logical regulatory networks based on steady-state experimental data. Our algorithm is based on DBNs and computes the expected value of the nodes by including an algebraic interpretation of the logical gates. The FALCON pipeline is shown in Figure 1.

2 Methods

2.1 Modelling of logical networks

FALCON models biological regulatory systems as DBNs, which are directed graphical models defined by the set of n nodes with $X = [0,1]^n$ and the probability distribution $P(X_t|X_{t-1}) = \prod_{i=1}^{n} P(X_t^{(i)}|Pa(X_t^{(i)}))$ where $X_t^{(i)}$denotes the i'th node at time t and $Pa(X_t^{(i)})$ represents the parents of $X_t^{(i)}$. These conditional probabilities are implicitly formulated by the structure of the network. The different nodes represent the different molecules of the system, with a value corresponding to the degree to which these molecules exist in their active form (for example, phosphorylated proteins). These node values can be understood as the proportion of the molecules in the system being active, or as the probability for a randomly chosen molecule to be active at time t.

In the FALCON framework, each molecular interaction is formulated as a logical predicate associated with a weight quantifying the relative importance of that specific interaction. We model different types of biochemical interactions with two types of edges: positive and negative edges connect activators and inhibitors to their downstream targets. Hyperedges corresponding to the '*AND'* and '*OR'* logical operations link multiple nodes to an output node, and model the activity of protein complexes and competition, respectively. Each edge and hyperedge is associated to a weight $k_j^{(i)}$ representing the relative influence of the upstream node to the downstream node. Because our modelling framework is grounded in Bayesian theory, the weights need to obey the law of total probability: for each node $X^{(i)}$ having a set $j_+$of m activating functions, we ensure the sum of activating weights $\sum_{j_+=1}^{m} k_{j_+}^{(i)} = 1$. Similarly, as weights of inhibiting interactions materialize the relative inhibition of upstream nodes, for nodes having a set j_- of l inhibiting functions, we ensure that $0 \leq \sum_{j_-=1}^{l} k_{j_-}^{(i)} \leq 1$.

Given a network structure established from prior knowledge, a set of parameters (weights) and a set of experimental conditions, the steady-state of the network is computed for each of the conditions and the values of the nodes corresponding to the measured species are recorded. For each one of the conditions, the nodes of the network are initialized with random values, except for the nodes considered as inputs (external to the system) for which the value is determined by the experimental condition and kept constant. The network is then updated repeatedly by computing synchronously for each node the expected value of its probability distribution, given the value of its parent nodes and the weights

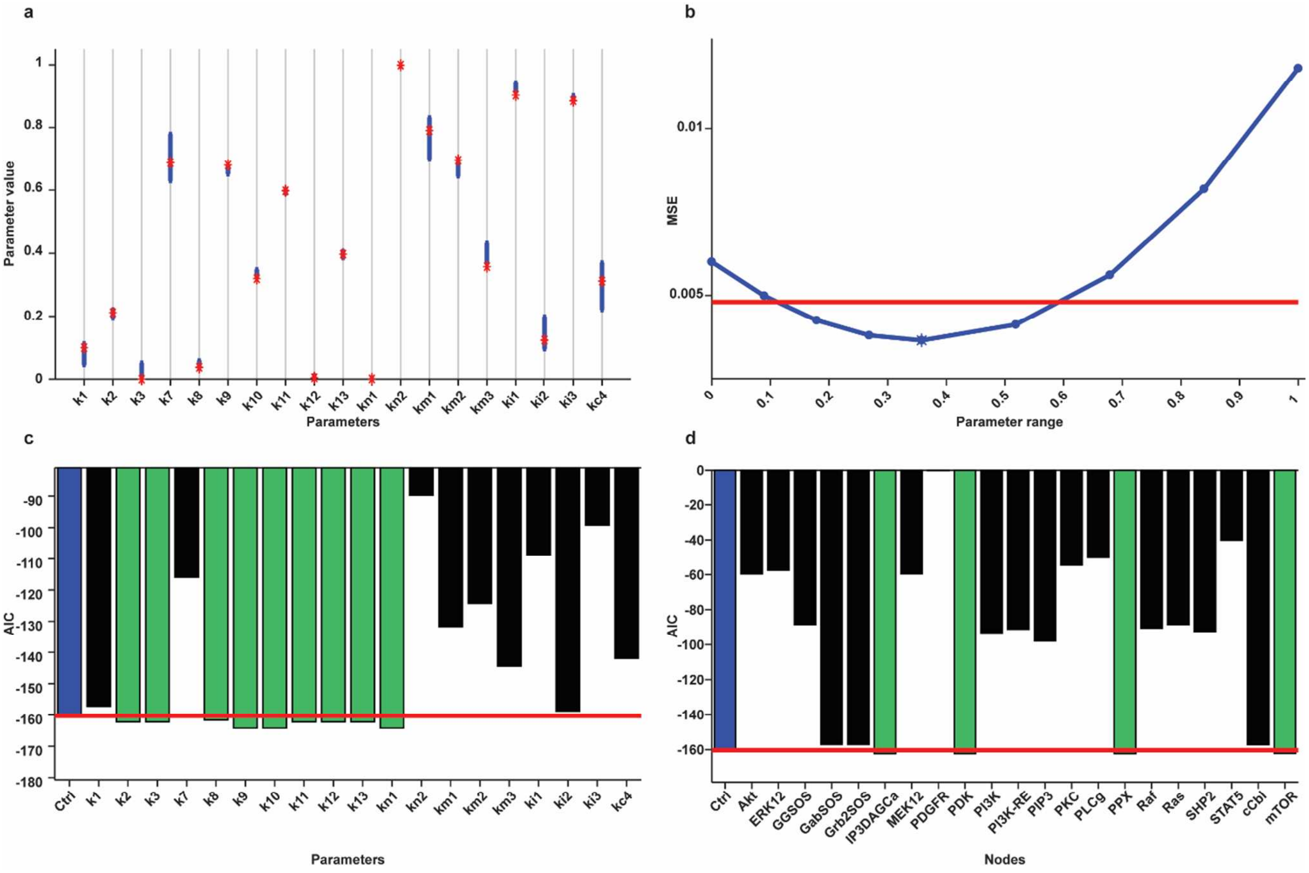

Figure 2. Analyses of optimized model in FALCON (PDGF model). a: Parameter robustness analysis ; red stars: optimal parameter values, blue bars: standard deviations of parameter values fitting to 10 resampling datasets. b: Parameter identifiability analysis of parameter 'km3' from panel a; Red line: threshold used to speed up computations in the 'fast' mode. c: Interaction knock-out analysis. d: Node knock-out analysis. In panels c and d, the color of the bars indicates the sign of the difference with the base model (blue). Green indicate better models (AIC_{model}<AIC_{base}), black indicates worse ones. Abbreviations: MSE = mean squared error, AIC = Akaike Information Criterion.
]

associated with each interaction.

$$X_t^{(i)} = \sum_{j_+=1}^{m} k_{j_+}^{(i)} Pa(X^{(i)})_{t-1}^{(j_+)} * (1 - \sum_{j_-=1}^{l} k_{j_-}^{(i)}\ Pa(X^{(i)})_{t-1}^{(j_-)})$$

Because all nodes at each update are considered as independent, the inputs values of '*AND'* logical gates are multiplied. The computation of '*OR'* gates follows De Morgan's law, i.e. the complement of the union of two sets is the same as the intersection of their complements. Inputs pointing to the same child node that are not members of a logical gate are summed. Table 1 summarizes the different types of interactions explicitly formulated in our framework. The algebraic formulas used for the computations can be directly derived from the conditional probability tables of the DBN formulation of the logical interactions.

Table 1: Different types of biological interactions modelled by different Boolean functions and their algebraic representations.

Biological equivalent	Graphical form	Algebraic computation
Activation	A → Z (k)	$Z_{t+1} = A_t * k$
Inhibition	A -| Z (k)	$Z_{t+1} = 1 - (A_t * k)$
Complex formation	A *AND* B → Z (k)	$Z_{t+1} = A_t * B_t * k$
Competitive interaction	A *OR* B → Z (k)	$Z_{t+1} = 1 - [\ (1\text{-}A_t) * (1\text{-}B_t) * k]$
Non-competitive interaction	A → Z (k_1) B → Z (k_2)	$Z_{t+1} = A_t * k_1 + B_t * k_2$ (with $k_1 + k_2 = 1$)

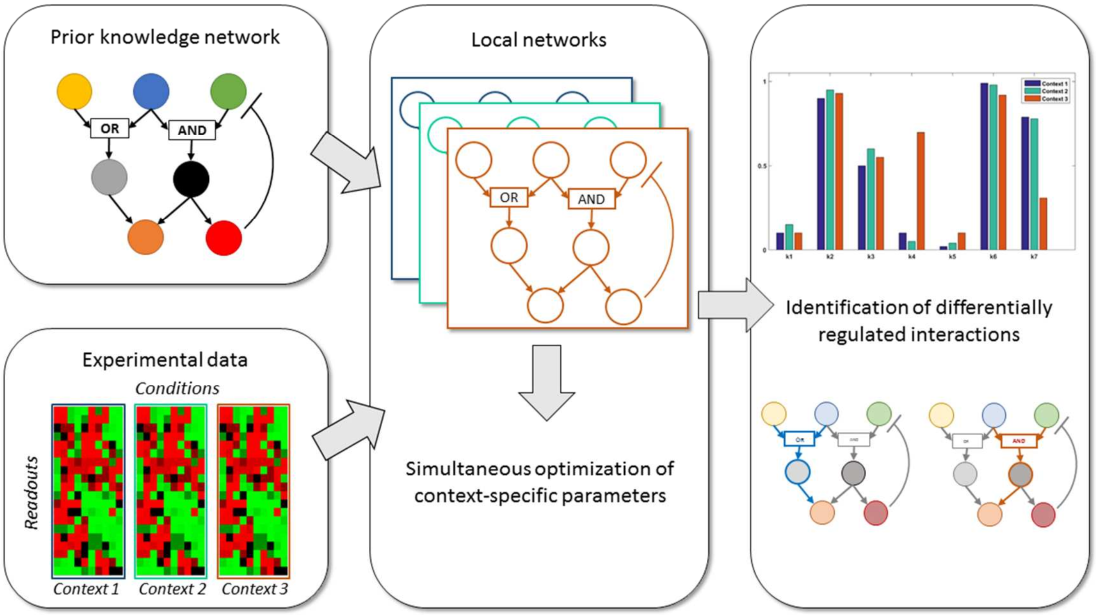

Figure 3. Differential analyses in FALCON. The same prior knowledge model is contextualized in parallel with different datasets corresponding to different contexts. Subsequent analysis can identify context-specific parametrizations and topologies.

The resulting dynamical system converges to a steady-state where each node value corresponds to the normalized equilibrium concentration of the activated form of the molecule in the system.

2.2 Contextualization algorithm

Objective function. To perform the contextualization of the model with experimental data, we extract from the network at steady-state the value of the nodes corresponding to the measurements, compare them with the normalized values from the experimental data and compute the mean squared error (MSE) between the estimated values and the measurements. We minimize this measure of the error by optimizing the value of the weights using a gradient-descent algorithm. To guarantee high efficiency while allowing for arbitrary degrees of recurrence in the networks, we use the interior-point method (Waltz *et al.*, 2004). A scheme of the FALCON workflow is presented in Fig. 1.

Rapid Optimisation. Using the gradient-descent optimization algorithm fmincon with interior-point method, FALCON is able to rapidly estimate the set of weights that minimizes the objective function. Random initialization of the weights is done either from a uniform distribution across the [0, 1] range, or from a truncated normal distribution centred on 0.5, depending on users' choice. Normally distributed initial values have been shown to improve learning for deep neural networks (Glorot and Bengio, 2010) and in our hands, increase the speed of convergence of the optimisation algorithm.

2.3 Subsequent analyses on optimized logical networks

Once a set of parameters has been inferred from a given topology and dataset, a series of additional analyses can be performed to gain more insight into the systems-level properties of the regulatory network being modelled as summarised in Figure 2.

Robustness of optimised parameter values. Depending on the topology of the network, the uncertainty in the measurement of some nodes can have more impact on the parameter values of the model than others. FALCON can analyse the uncertainty on inferred parameter values by sampling a user-defined number of artificial datasets based on original experimental measurements and determining the weights of the model in the light of the new data (Fig. 2a). The artificial datasets are constructed from the average experimental measurements and their associated error, assuming normally-distributed residuals.

Identifiability analysis. In order to assess the identifiability of the model parameters, an approach similar to Raue *et al.* is applied (Raue *et al.*, 2009). For each parameter, the algorithm samples the range of possible parameter values [0, 1], and re-optimizes the model under the additional constraint of this parameter being fixed to each one of the sampled values. In order to obtain the most meaningful results we sample the same number of points on both sides of the optimal value. We include the option to skip the most extreme values based on a threshold determined by the resampling analysis (red line, Fig. 2b), thereby accelerating computations. The resulting MSE profiles allow to determine which parameters are well constrained by the experimental measurements.

Interactions Knockouts. FALCON allows the systematic removal of each edge in the network and provides a graphical output showing the effect on the global fitness of the model. The models are compared using the Akaike Information Criterion (Burnham and Anderson, 2004), which balances goodness-of-fit with model complexity (Fig. 2c). By using this additional analysis, it is possible to differentiate the crucial edges of the system from the ones that are dispensable, which can be pruned out.

Nodes Knockouts. A frequent goal of systems biology analyses is to identify the crucial molecules of a regulatory network. Often performed via network topological properties (centrality measures), this identification is of particular interest in the case of target discovery efforts. FALCON allows the systematic evaluation of models in which each node is removed from the network. The comparison of these models using the Akaike Information Criterion allows to identify these crucial nodes not only from topological properties but from the effect their removal has on the behaviour of the entire system (Fig 2d).

Differential regulation. In many real-life modelling applications, a system is studied in different contexts. For example, during a drug screen, the same signalling pathways are studied for different cell lines, or over time. One goal of systems biology is to identify differences between the contexts in the way the system is regulated. FALCON automates such analyses by optimizing identical models in parallel for multiple series of experimental conditions. Users can discover which parts of the network are activated or shut down between cell lines/time points, and this may lead to the identification of specific interventions strategies for each context (Fig. 3).

3 Pipeline and Performance

FALCON is a highly efficient optimisation tool that is capable of contextualizing small-to-large biological networks. For an easy input of model structure and experimental data, FALCON accepts different file formats (.txt, .xls, .xlsx, .csv) which are subsequently used to build logical models. Inference of network structure, interaction matrices, and parameter constraints are fully automated, and the toolbox outputs a user-friendly summary comprising the optimized weights for the different interactions, both in text and graphical forms. To facilitate the use of our toolbox, we included a graphical user interface (GUI) to guide users through the different steps of the workflow. Users who are more comfortable with the *MATLAB* language can instead choose to use the provided driver script for full flexibility.

To showcase the performance of our toolbox, we provide four examples, including the replication of several studies, each presenting a particular challenge for the toolbox. The results of our tests are shown in Table 2. All computations were performed on a desktop PC with 16 GB RAM and an Intel® Xeon® CPU E3-1246 v3, 3.5 GHz with Matlab 2016b.

Toy model: we demonstrate the basic functionality of FALCON on a 6-node toy model, comprising both positive and negative interactions, as well as a Boolean AND gate. The structure of this network, associated synthetic data and trained model are illustrated in Figure S1 in the Supplementary Material.

PDGF: we used FALCON to optimise a platelet-derived growth factor signalling model (Trairatphisan *et al.*, 2016), comprising 30 nodes and 37 interactions (19 free parameters). The dataset was assembled from the quantification of 6 proteins by western blot analysis in HEK293 cells expressing a constitutively active form of the PDGF receptor, in the presence or absence of two types of perturbations: single-point mutations of tyrosine residues on the PDGF receptor associated with the recruitment sites of downstream signalling molecules, and kinase inhibitors. We obtained a fitting cost (MSE=0.0041) and parameter values very similar to the original study, where the tool optPBN (Trairatphisan *et al.*, 2014) was used to perform the optimization, and in accordance with it, we are able to train the network with single perturbations and accurately predict the signalling profiles of combined perturbations experiments (see Supplementary Materials).

Apoptosis: we replicated a modified model of a previous study in which a large Boolean model of apoptosis was used to investigate non-linear dose-effects of UV radiation on cultured hepatocytes (Schlatter *et al.*, 2009; Trairatphisan *et al.*, 2014). The model comprises 138 nodes and 160 interactions (41 free parameters). We correctly estimated apoptosis levels and the other associated experimental measures, and could draw the same conclusions as the original study concerning the importance of cross-talks, especially between Caspase 8 and NFKB (see Supplementary Materials). While the original study used the software CellNetAnalyzer (Klamt *et al.*, 2007), which uses a multi-value Boolean formalism and concentrates on network properties, a previous replication with the optPBN toolbox (Trairatphisan *et al.*, 2014) could infer more quantitative properties, but at the expense of long computation times. Analysis of this network and data with FALCON is comparatively very fast with up to 170-fold improvement (FALCON: 76 seconds; optPBN: 4 hours 40 minutes) and we obtained a fitting cost (FALCON: MSE=0.017) comparable with the previous studies (optPBN: MSE=0.011; Schlatter *et al.*: MSE=0.013). In comparison, CellNetAnalyzer, using discrete Boolean modelling and only able to consider either full activation of complete inactivity of the molecules, achieves a worse fit (MSE: 0.056). The comparison of the inferred molecular states of optPBN and FALCON can be found in Supplementary Figure S6.

MAPK: we compared the performance of our tool with the software CellNOptR (Terfve *et al.*, 2012; MacNamara *et al.*, 2012) in the fuzzy logic mode (CNORfuzzy) for quantitative optimisation of model states. Using the toy example provided, which is the optimized network of the DREAM4 challenge and contains 22 nodes, 36 interactions and 25 experimental conditions (Prill *et al.*, 2011), we obtained a similar fitting cost with FALCON (MSE=0.036) and with CellNOptR (MSE=0.032) but with a gain of speed of about 44 times (see Table 2).

Table 2. Accuracy and computation times for the different examples. The cost is expressed as MSE (mean squared error) and the speed is expressed in seconds (s).

Example	Nodes	Edges / Parameters	Datapoints	Cost	Speed
Toy (artificial)	6	3 / 3	10	0	< 1s
PDGF	30	19 / 19	36	0.004	1.3s
Apoptosis	138	160 / 41	18	0.017	76s
MAPK [FALCON]	22	32 / 32	175	0.036	1.1s
MAPK [CNORfuzzy]	22	32 / 92	175	0.032	47.4s

4 Discussion

We present FALCON as an alternative tool for the efficient optimization and comprehensive analysis of logical models of regulatory networks. Our modelling framework, based on DBNs, is able to determine qualitative and quantitative features of the systems being modelled. Node values, being comprised in the interval [0, 1], represent the probabilities for molecules to be in their active state at equilibrium. They can also be understood as the normalized average activities of the nodes. The computed parameters, or weights, also comprised in the interval [0, 1] and subject to the law of total probability, represent the probabilities for the designated interactions to influence downstream nodes. They can also be interpreted as the relative influences of the parent nodes on their children nodes and are useful in assessing the flow of the signal transduction.

FALCON, through its GUI, is easy to use for scientists without extensive modelling experience. FALCON is also very fast compared to similar tools based on PBNs, and surpassed CellNOptR in our test. The low computation costs make it possible to analyse the models at the systems level through a series of bundled additional analyses which allow to answer a number of biologically important questions: whether the parameter values are well constrained by the available data, how the experimental error influences the confidence in the parameter values, and which are the nodes and interactions most crucial to the behaviour of the system versus the ones that can be pruned out. Together, our results suggest that FALCON is a very useful software for rapid model exploration, especially for large networks and large datasets.

Compared to the popular package CellNOptR, the FALCON pipeline is faster in contextualizing a small graphical model with quantitative data. The inferred parameters are also more intuitively understandable as the relative strength of the interactions, while CellNOptR combines linear and Hill's equations in a way that does not encourage direct interpretation. This relative complex formulation, together with the multiple concurrent formalisms proposed and the increased computational cost suggest reserving this tool for more complex tasks, while FALCON is better adapted for exploratory studies of larger networks and datasets.

Future development of the FALCON toolbox will include full compatibility with established model representation formats (SBML-Qual, BioPAX), and the conversion of the toolbox to other languages, like R, Python and C++. One particular aspect that we regard as highly interesting is the use of FALCON to explore model topologies in a large-scale, systematic way to uncover previously unknown mechanisms in regulatory networks.

In terms of applications, we demonstrated that FALCON is applicable to model signal transduction networks and could easily be extended to study other biological regulatory systems. We envision that FALCON has the potential to be widely adopted by the computational biology community, including biologists with limited programming experience.

Acknowledgements

We would like to acknowledge Dr. Jun Pang, Dr. Andrzej Mizera, Prof. Dr. Dagmar Kulms, Greta del Mistro and Dr. Thomas Pfau for the fruitful discussions and suggestions on the modelling and analytical pipelines.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642295 (MEL-PLEX) and the Luxembourg National Research Fund (FNR) within the projects MelanomSensitivity (BMBF/BM/7643621) and ALgoReCEll (INTER/ANR/15/11191283).

Conflict of Interest: none declared.

References

Bansal M., Belcastro V., Ambesi-Impiobato A. and di Bernardo D. (2007), How to infer gene networks from expression profiles. Molecular Systems Biology, 3(78), 1-10.

Burnham, K. P. and Anderson, R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociological Methods & Research, 33(2), 261–304.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS), 9, 249-256.

Kauffman, S. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of theoretical biology, 22(3), 437–467.

Klamt S., Saez-Rodriguez J. and Gilles E. D. (2007) Structural and functional analysis of cellular networks with CellNetAnalyzer. BMC Sys. Biol. 1:2.

Lähdesmäki, H., Hautaniemi, S., Shmulevich, I., and Yli-Harja, O. (2006). Relationships between probabilistic Boolean networks and dynamic Bayesian networks as models of gene regulatory networks. Signal processing, 86(4), 814–834.

Le Novère, N. (2015). Quantitative and logic modelling of molecular and gene networks. Nature Reviews Genetics 16(3), 146-158.

Li, F., Long, T., Lu, Y., Ouang, Q., and Tang, C. (2004). The yeast cell-cycle network is robustly designed, PNAS 101, 4791-4786.

Mizera, A., Pang, J., and Yuan, Q. (2016). ASSA-PBN 2.0: A Software Tool for Probabilistic Boolean Networks. In Computational Methods in Systems Biology, pages 309–315.

Morris, M. K., Saez-Rodriguez, J., Sorger, P. K., and Lauffenburger, D. A. (2010). Logic-based models for the analysis of cell signaling networks. Biochemistry, 49(15), 3216–3224.

Prill, R.J., Saez-Rodriguez, J., Alexopoulos, L.G., Sorger, P.K., et al.. (2011) Crowdsourcing network inference: The DREAM Predictive Signaling Network Challenge. Sci. Signal. 4(189) mr7.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics, 25(15), 1923–1929.

Saez-Rodriguez J, Simeoni L, Lindquist JA,Hemenway R, Bommhardt U, Arndt B, Haus UU, Weismantel R, Gilles ED, Klamt S, et al. (2007). A logical model provides insights into T cell receptor signaling. PLoS Comput Biol, 3:e163.

Schlatter,R. et al. (2009) ON/OFF and beyond - a Boolean model of apoptosis. PLoS Comput. Biol., 5, e1000595.

Terfve, C., Cokelaer, T., Henriques, D., MacNamara, A., Goncalves, E., Morris, M. K., van Iersel, M., Lauffenburger, D. a., and Saez-Rodriguez, J. (2012). Cell- NOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. BMC Systems Biology, 6, 133.

Trairatphisan, P., Mizera, A., Pang, J., Tantar, A. A., Schneider, J., and Sauter, T. (2013). Recent development and biomedical applications of probabilistic Boolean networks. Cell Communication and Signaling : CCS, 11(46).

Trairatphisan, P., Mizera, A., Pang, J., Tantar, A. A., and Sauter, T. (2014). optPBN: An optimisation toolbox for probabilistic Boolean networks. PLoS ONE, 9(7).

Trairatphisan, P., Wiesinger, M., Bahlawane, C., Haan, S., and Sauter, T. (2016). A Probabilistic Boolean Network Approach for the Analysis of Cancer-Specific Signalling: PDGF Signalling in GIST. PLoS One, 11(5), e0156223.

Villaverde, A. F. and Banga, J. R. (2013). Reverse engineering and identification in systems biology: strategies, perspectives and challenges. Journal of The Royal Society Interface.

Waltz, R. A., Morales, J. L., Nocedal, J., and Orban, D. (2004). An Interior Algorithm for Nonlinear Optimization That Combines Line Search and Trust Region Steps. Mathematical Programming, 107(3), 1–20.