Bioinformatics, 35(12), 2019, 2159–2161
doi: 10.1093/bioinformatics/bty916
Advance Access Publication Date: 15 November 2018
Applications Note

OXFORD

Gene expression

GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks

Thomas Moerman[1,2,*], Sara Aibar Santos[3,4], Carmen Bravo González-Blas[3,4], Jaak Simm[2,5], Yves Moreau[2,5], Jan Aerts[1,2] and Stein Aerts[3,4,*]

[1]KU Leuven ESAT/STADIUS, VDA-lab, [2]IMEC Smart Applications and Innovation Services, [3]Laboratory of Computational Biology, VIB Center for Brain & Disease Research, [4]Department of Human Genetics, KU Leuven, Leuven, Belgium and [5]KU Leuven ESAT/STADIUS, Bioinformatics Lab

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on February 23, 2018; revised on September 5, 2018; editorial decision on October 10, 2018; accepted on November 14, 2018

Abstract

Summary: Inferring a Gene Regulatory Network (GRN) from gene expression data is a computationally expensive task, exacerbated by increasing data sizes due to advances in high-throughput gene profiling technology, such as single-cell RNA-seq. To equip researchers with a toolset to infer GRNs from large expression datasets, we propose GRNBoost2 and the Arboreto framework. GRNBoost2 is an efficient algorithm for regulatory network inference using gradient boosting, based on the GENIE3 architecture. Arboreto is a computational framework that scales up GRN inference algorithms complying with this architecture. Arboreto includes both GRNBoost2 and an improved implementation of GENIE3, as a user-friendly open source Python package.

Availability and implementation: Arboreto is available under the 3-Clause BSD license at http://arboreto.readthedocs.io.

Contact: thomas.moerman@esat.kuleuven.be or stein.aerts@kuleuven.vib.be

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

1 Introduction

Unravelling the regulatory programs that produce a given gene expression profile has long been one of the major challenges in genomics. Using statistical or machine learning techniques, it is possible to uncover co-expression patterns among regulators and target genes, and infer potential regulatory associations that are represented in Gene Regulatory Networks (GRNs). Current single-cell RNA-seq (scRNA-seq) technologies (e.g. Drop-seq, Macosko *et al.*, 2015)—that are able to determine the transcriptional profile of tens of thousands of individual cells—have opened unprecedented opportunities to improve the quality and resolution of these networks. However, the increase in dataset size poses the new challenge of computational time efficiency. Our work aims at providing an efficient and scalable toolset that addresses the problem of inferring high quality GRNs from such large datasets in a reasonable amount of time.

Our work is inspired by GENIE3 (Huynh-Thu *et al.*, 2010), winner of the DREAM5 network challenge. The remarkable simplicity of GENIE3's GRN inference architecture makes it highly parallelizable. In previous work, we exploited this property and developed GRNBoost (Aibar *et al.*, 2017), a proof-of-concept Apache Spark pipeline for scalable GRN inference, using Gradient Boosting Machine (GBM) regression (Friedman, 2001). Building on that experience, we here propose two results that combinedly form a user-friendly and performant system for GRN inference. GRNBoost2 is a GBM-based GRN inference algorithm that focuses on efficiency while achieving excellent scores on the DREAM5 network benchmark. Arboreto is a computational framework that scales the workload of GRN inference algorithms complying with the GENIE3 architecture. We address multiple shortcomings of the GRNBoost proof-of-concept: GRNBoost2 introduces a self-tuning mechanism that replaces the use of a global estimate for the number of decision trees in the boosting ensembles. Arboreto performs workload scheduling dynamically rather than statically and is designed as a generic framework. It currently offers GRNBoost2, a reimplemented version of GENIE3 (referred to as: GENIE3/A) and offers the opportunity to accommodate future GRN inference algorithms.

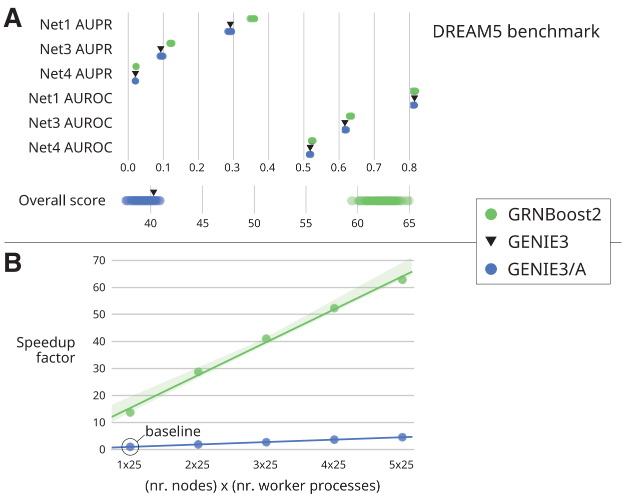

Fig. 1. (**A**) DREAM5 benchmark scores of 100 runs with GRNBoost2 and GENIE3/A. Black triangle markers depict the reported GENIE3 scores (Marbach *et al.*, 2012). (**B**) Speedup factor of GRNBoost2 and GENIE3/A with respect to the single node GENIE3/A baseline

2 Materials and methods

Like GENIE3, GRNBoost2 belongs to the class of *regression-based* GRN inference methods (Huynh-Thu and Sanguinetti, 2018). For each gene in the dataset, a tree-based regression model is trained to predict its expression profile using the expression values of a set of candidate transcription factors (TFs). Each model produces a partial GRN with regulatory associations from the best predicting TFs to the target gene. All regulatory associations are combined and sorted by importance to finalize the GRN output.

Gradient boosting is a machine learning method that builds regression or classification models by additively combining weak learners, typically shallow decision trees. Previous results (Sławek and Arodź, 2013) demonstrated that GBM regression models are suitable for predictor selection in a GRN inference context. GRNBoost2 employs a regularized stochastic variation on GBMs. It equips GBM regressions with a heuristic *early-stopping* regularization strategy using out-of-bag improvement estimates. Each new decision tree is trained in function of a random subset of observations (90%, hence stochastic), whereas the remaining (10%, out-of-bag) observations are used to calculate an estimate of the loss function improvement entailed by adding that tree to the ensemble. When the average of the last n improvement values drops below 0, the early-stopping criterion is met and no more trees are added to the ensemble. This self-tuning mechanism ensures that each regression model contains 'just enough' trees with respect to the learning rate η. Regressions that do not display net improvement early on are aborted and thus prevented from causing useless computational workload. The GRNBoost2 hyperparameters are discussed in Supplementary Section S2.

The regression-per-target GRN inference architecture consists of a large number of computations that can be performed independently, which makes them highly parallelizable. As GRNBoost2 and GENIE3 have the same abstract input/output signature, we isolated them as plugin functions from the framework that organizes the computational workload, called Arboreto. Arboreto (Supplementary Fig. S1) is implemented using Dask (Rocklin, 2015), a parallel computing library for the Python programming language. With Dask, a computation is specified as a directed graph of tasks with data dependencies and executed using a Dask scheduler. The scheduler delegates the tasks in the graph to worker processes running on one or multiple compute nodes connected by a network. The computation graph is submitted to the workers in two steps: the predictor matrix is *broadcasted* to the different workers, the regression instances are distributed evenly between workers. The scheduler operates dynamically, reassigning tasks from busy to idle workers when necessary. Upon completion of the graph, all partial networks are aggregated into a list of weighted regulatory links, sorted by importance and thresholded.

We assessed GRN inference quality of GRNBoost2 and GENIE3/A using the DREAM5 benchmark. As both algorithms are stochastic, we ran them each 100 times with different initialization seeds. Figure 1A illustrates that both algorithms yield high quality results irrespective of the random initialization seed. The inference quality results of GENIE3/A are consistent with the original GENIE3 implementation. To measure the speedup using distributed hardware, we inferred the GRN from a scRNA-seq dataset of 40 000 cells and 12 953 genes (derived from Macosko *et al.*, 2015) using clusters of respectively 1–5 compute nodes (dual Intel Xeon E5-2680 v3 CPUs, 512GB RAM), running 25 Dask worker processes each. Figure 1B illustrates the speedup factors of GRNBoost2 and GENIE3/A with respect to the baseline of single-node GENIE3/A. The GRNBoost2 speedup on the scRNA-seq dataset is achieved by two factors. Firstly, the bias-reducing effect of gradient boosting affords the use of shallower decision trees than random forest. Secondly, using early-stopping GRNBoost2 constructed over 80% fewer decision trees in total than GENIE3/A. The benefit of early-stopping is proportional to the number of gene profiles with few or noisy expression levels. As expected, the Arboreto framework scales both GRN inference algorithms approximately linearly with respect to compute resources.

While many approaches for GRN inference have been proposed (see Huynh-Thu and Sanguinetti, 2018 for an overview), our work distinctively offers an efficient and scalable implementation that aims at bridging the gap between powerful GRN inference methods and increasingly large datasets.

Acknowledgements

The authors thank Geurts, P. and Huynh-Thu, V.A. for the GENIE3 code and helpful discussions, and thank Spanier, K., Kalender Atak, Z., Thrupp, N., De Waegeneer, M. and Van de Sande, B. for testing the software.

Funding

IMEC Strategic Funding, IWT Exaptation, FWO 06260, Special Research Fund (BOF) KU Leuven (PF/10/016 to SA, YM and JA), ERC Consolidator Grant (724226_cis-CONTROL to SA) and Foundation Against Cancer (2016-070 to SA).

Conflict of Interest: none declared.

References

Aibar,S. *et al.* (2017) SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*, **14**, 1083–1086.

Friedman,J.H. (2001) Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, **29**, 1189–1232.

Huynh-Thu,V.A. *et al.* (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**, e12776.

Huynh-Thu,V.A. and Sanguinetti,G. (2018) Gene regulatory network inference: an introductory survey. *arXiv: 1801.04087*.

Macosko,E.Z. *et al*. (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.

Marbach,D. *et al*. (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**, 796–804.

Rocklin,M. (2015) Dask: parallel computation with blocked algorithms and task scheduling. In: Kathryn,H. and James,B. (eds) *Proceedings of the 14th Python in Science Conference*, matthew_rocklin-proc-scipy-2015, pp. 130–136.

Sławek,J. and Arodź,T. (2013) ENNET: inferring large gene regulatory networks from expression data using gradient boosting. *BMC Syst. Biol*., **7**, 106.