Journal Name

ARTICLE

Dissection of the Module Networks Implementation "LemonTree": Enhancements towards Applications in Metagenomics and Translation in Autoimmune Maladies

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Youtao Lu, [a,b] ‡ Xiaoyuan Zhou,[a] ‡ and Christine Nardini*[c,d,a]

Under the current deluge of omics, module networks distinctively emerge as methods capable not only to identify inherently coherent groups (modules), thus reducing dimensionality, but also to hypothesize cause-effect relationships between modules and their regulators. Module networks were first designed in the transcriptomic era and further exploited in the multi-omic context to assess (for example) miRNAs' regulation on genes expression. Despite a number of available implementations, expansion of module networks to other omics is constrained by a limited characterization of the solutions' (modules plus regulators) accuracy and stability -- an immediate need for the better characterization of molecular biology complexity *in silico*. We hence carefully assessed for LemonTree -a popular and open source module networks implementation- the dependency of the software performances (sensitivity, specificity, false discovery rate, solutions' stability) on the input parameters and on the data quality (sample size, expression noise) based on synthetic and real data. In the process, we uncovered and fixed an issue in the code for the regulators' assignment procedure. We concluded this evaluation with a table of recommended parameter settings. Finally, we applied these recommended settings to gut-intestinal metagenomic data from rheumatoid arthritis patients, to characterize the evolution of the gut-intestinal microbiome under different pharmaceutical regimens (methotrexate and prednisone) and we inferred innovative clinical recommendations with therapeutic potential, based on the computed module network.

Introduction

The usage of module networks for the identification of causal relationships between gene clusters (proxies for genes functions) and their molecular regulators (specific to the experimental conditions) was first applied to yeast as an extension of Bayesian networks (BN), assuming that variables with similar statistical behaviors (namely, transcriptional profiles with similar distribution and tight correlation) can be grouped into *modules*, where variables (transcripts) share the same regulators (for example transcription factors).[1] This translates in statistical terms as a common conditional probability distribution, and models the fact that, in a large domain like a molecular pathway, genes involved in a specific function tend to share transcription factor binding sites and other types of regulators and to express in a coordinated manner.[2, 3] BN are extremely powerful and adapted to model biological interactions (networks) owing to their capacity to learn from incomplete and noisy data and to hypothesize causal relationships.[4] In particular, although direct feedback is not modellable with BN, the notion of feedback can be expressed by indirect feedback, which, despite not modelling the exact biological phenomenon, achieves globally the needed signal transmission and allow appropriate representation of the final (steady) state of the modelled pathway. Module networks infer a combinatorial (hierarchical) cascade of regulators called *regulatory program* (modelling for example a pathway), constructed in the form of a *regression tree* where leaves contain the experiments (samples) grouped by similarity and nodes represent the corresponding regulators. Leaves and nodes are computed in two steps: the *gene clustering* step and the *regulators assignment* step. In the first step, clustering solutions from multiple instantiations are collapsed into *tight* clusters (maximum between-cluster and minimum within-cluster distance); in the second step, module-wise, regulators are iteratively evaluated and assigned to a particular node (i.e. *split*), according to the fitness of their expression profiles (up/down or down/up transition across different experimental conditions) to the responses of genes regulated by this node. The two steps (*gene clustering* and *regulators assignment*) are computed independently.

Module networks algorithms have been implemented in a variety of software packages, such as Genomica,[5] LeMoNe (Learning Module Networks) and MERLIN.[6-12] Among them,

[a] CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, P.R.China; University of Chinese Academy of Sciences, Beijing 100049, P.R. China
[b] Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA
[c] CNR IAC "Mauro Picone" Via dei Taurini 19, 00185 - Roma, Italy
[d] Personal Genomics S.r.l, Strada Le Grazie 15-37134 Verona, Italy. Email: christine.nardini.rsrc@gmail.com
†Electronic Supplementary Information (ESI) available: Data S1-S5. See DOI: 10.1039/x0xx00000x
‡These authors contributed equally to this work.

LeMoNe has been under development for several years, and its most recent version "LemonTree", used in this analysis, has been majorly improved in user friendliness and integration of the pipeline (written in pure Java, no longer dependent on Perl and Matlab to process intermediate results).[6] The major distinctive characteristic of LemonTree is the extensive usage of the "ensemble approach" – a machine learning strategy that combines multiple models to obtain better performances than single model approaches and allows flexibility in the final structure.[13] The ensemble approach is used both in the *gene clustering* and in the *regulators assignment* step where a stochastic sampling technique confines the regulator list to the most likely ones.

To perform our tests we first exploit the synthetic data software SynTReN [14], for the analysis of the solutions' accuracy, using metrics including sensitivity, specificity and false discovery rate (vital to practical applications of the software) under different parameter settings and data quality. We then deepen this analysis with quantification of the solutions' stability on a real cancer dataset,[15] and implement a correction to the original code, which overcomes an issue compromising the *regulators assignment* task when candidate regulators have opposite expressions between a partition of experiments (samples).

The emerging guidelines are applied in a novel inferential setting: the changes in the composition of the gut intestinal (GI) microbiome in the progression of rheumatoid arthritis (RA).

Materials and Methods

Real data for performance evaluation

Real data contain paired mRNA and miRNA expression profiles from 89 cancer samples.[15] We preprocessed the mRNA expression data to be zero centered with standard deviation SD = 1 for each gene, and we filtered low dispersion (SD < 0.5) following LemonTree specifics (3826 mRNAs were removed). The candidate regulator list consists of 2060 miRNAs and mRNAs annotated to functions GO:0007165 (signal transducer activity) and GO:0003677 (DNA binding).[7] We subsampled this list to derive four candidate regulator lists at different lengths defined by parameter: *reg* = 10, 50, 100, 500. For more details see Table 1.

Synthetic data for performance evaluation

To warrant objective evaluations of the performance, SynTReN was used to generate a series of 200-gene expression datasets.[14] SynTReN subsamples two networks from the yeast regulatory gene network: the *foreground* network, consisting of true yeast regulators and their regulated genes and the *background* network, using false candidate regulators.[16] To simulate interactions, correlations were imposed between genes across different treatments/conditions in the *foreground* but not in the *background*. Noise was simulated according to two factors: biological noise which accounts for stochastic variations in gene transcription, and technical noise, approximated by a lognormal distribution.[14] For the sake of simplicity, we varied biological noise (*bn*) from level 0.1 to 0.5 while we kept technical noise constant, to reflect experimental conditions that usually exploit only one technology (microarray, sequencing, etc.). After adding noise to both networks, expression data were sampled from each node and finally exported in a matricial format (rows for genes, columns for samples). Given the small number of true regulators (17), we kept the candidate regulator list unvaried to avoid loss of information due to random subsampling. Additionally, we simulated a variable number of replicates (*nreplicate*) to reflect different sample sizes.

We then performed quality check on the simulated data, to ensure that noise levels and expression correlations in the *foreground* and *background* follow SynTReN's parameter specification. Due to the high coexpression levels in the *foreground* versus the *background,* we treated the *foreground* as a single module and the *background* as a source of false regulators. For more details see ESI,† Data S1.

Parameters and evaluation metrics

LemonTree version 3.0.2 allows users to control four input parameters: *nc* (number of clustering solutions before collapsing them into tight clusters), *reg* (total number of candidate regulators), *nreg* (number of regulators assigned to a split), *perc* (percentage of top regulators output for each module, computed on the full list of assigned regulators). Performances are evaluated on the two outputs: modules and regulators.

Regarding modules, we assessed the stability using the number of tight clusters (NTC) by varying the input parameter *nc*.

Table 1 LemonTree input parameters and output performances.

Input Parameters	Tested Values	Data Type	Output Performances	
			Modules	Regulators
number of clustering solutions (*nc*)	*nc* = 5,20,30,50	real	NTC	-
number of candidate regulators (*reg*)	*reg* = 10,50,100,500	real		OAR
number of regulators assigned to each split (*nreg*)	*nreg* = 1,2,5,10,20,50,100,200	synthetic		AUC, TPR, FDR
	nreg = 5,10,20,50,100,200,500	real	-	OAR
top percentage of the full regulator list (*perc*)	*perc* = 1,5,10,20,…,100 %	synthetic		AUC, TPR, FDR
		real		OAR

The table lists LemonTree input parameters (Column 1) and the benchmark variables for output performances (Column 4). Column 2 lists the parameters' tested values, Column 3 the data source. For definitions of input parameters and output variables see Methods.

For regulators, we evaluated the *accuracy* with an array of metrics including true positive rate (TPR, sensitivity), false positive rate (FPR, 1-*specificity*), false discovery rate (FDR), as well as the corresponding receiver operating characteristic (ROC) curves and areas under the curve (AUC, *y*-axis TPR, *x*-axis FPR or FDR); *stability*, with the average overlapping of regulators (AOR) defined as $1/N \sum_{1 \le i,j \le N} |R_i \cap R_j| / |R_i|$, where *R* is a set of assigned regulators, i, j two different instantiations of the program and *N* is the number of instantiations. Regulators performances are assessed by varying the input parameters *reg*, *nreg*, *perc*. Definitions of all benchmarked parameters and metrics are summarized in Table 1.

Rheumatoid arthritis GI microbiome data

The 16S rRNA data from the GI tract were extracted from Scher *et al.* publication, collected from healthy individuals (HLT, 28 samples) and RA patients, either newly diagnosed (new onset RA, NORA, 39 samples) or chronically affected (CRA) and treated with disease-modifying anti-rheumatic drugs (DMARD, methotrexate [MTX], 9 samples) or corticosteroids (prednisone, 3 samples).[17] For all arms, nonsteroidal anti-inflammatory drugs (NSAIDs) were optional.

Although these are not longitudinal data, the new onset samples (NORA) characterize the early stage of the disease and the chronic samples (MTX, prednisone) characterize the later, treated stages of the disease. Therefore, the differentially abundant bacteria (comparison of NORA vs the HLT baseline) were assumed to model the initial state of the system (and as such to be candidate drivers -*regulators*- of the microbiome composition) whose final (i.e. chronic) state is represented by the composition observed in two different treatment arms (MTX and prednisone) where *modules* have been computed. Mathematically our assumption has the formalization $p_{k,R} = Pr(X_k^1 | \{x_r^0, r \in R\})$, where *k* denotes Module *k*, *R* the set of candidate regulators, x_r the abundance of regulator microbe *r*, *X* the partition of the regulated microbes' abundance (i.e. the regulatory program), and 0/1 denotes before/after treatments; $p_{k,R}$ is the posterior probability indicating the likelihood of the regulators *R* regulating Module *k*, and the algorithm aims at finding a set of regulators $R^* \in R$ that maximize the probability. As LemonTree is based on Bayesian networks that are acyclic and specifically we were inferring a causal relationship of how the presence or absence of the regulators influence the other ones, *candidate regulators* were excluded from the set of regulated microbes (i.e. in the modules) to forbid direct self-regulation, still allowing indirect regulation. Although self-regulation is a phenomenon that we miss to model, the microbiome community is known to rely highly on the metabolic and inflammatory environment contributed by bacteria of different genera working in synergy [18], which can be captured by this model.

Metagenomic and clinical data are available in ESI,† Data S2; for detailed preprocessing methods, input parameters and output modules/regulators see ESI,† Data S3.

Results and Discussion

Benchmarking results are organized in the first three subsections: the first concerns modules stability affected by *nc*, measured with NTC, using real data; the second characterizes the performances, reported using AUC on the regulators output for variations of the input parameters *nreg* and *perc*, with synthetic data (as biological noise -*bn*- and sample size -*nreplicate*- can be tuned). The third subsection assesses the stability of the regulators affected by *reg*, *nreg*, *perc*, measured with AOR using real data. It is noteworthy that the real data benchmark complements the synthetic for the accessibility to multiple modules and candidate lists, and is more adapted to the characterization of regulators stability.

At the end of these three testing subsections, we present motivation and results for LemonTree software correction. Finally, the last subsection describes the application, results and discussion of our parameter recommendations on GI microbiome data in RA.

Stability in module generation using real data

We first probed the dependency of the modules stability on the number of gene clustering solutions. We quantified the modules stability with the number of tight clusters. As shown in Fig. 1, after *nc* = 30, the median number of generated modules becomes stable, suggesting a saturation for the NTC that can be produced. Actually *nc* = 30 or *nc* = 50 have been proposed in LemonTree's publications.[7, 9] Our results hence confirm the authors' suggestions ($nc \geqslant 30$).

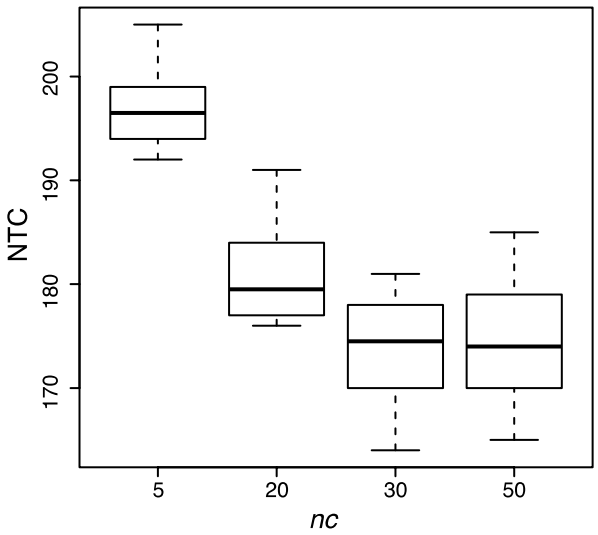

Fig. 1 Stability in module generation. *x*-axis: number of clustering solutions (*nc*); *y*-axis: number of tight clusters (NTC), each box summarizes NTC median and IQR , for 10 algorithms instantiations with increasing *nc*.

Regulators assignment using synthetic data

Because of the flexibility allowed by synthetic data, we report in this subsection several manipulations and the corresponding performances, assessed with a large number of performance metrics, to guarantee completeness of our results and robustness of the descending recommendation.

We first computed AUC (TPR-FPR curves) to give an overview of the performances on the synthetic data (Fig. 2). For most of the datasets and *nreg* values, LemonTree shows good performance in finding the true regulators, as AUCs are well above the randomness threshold of 0.5. In particular, we can inspect the trend as we vary the sample size (*nreplicate, rows*), the biological noise levels (*bn, columns*), and the number of candidate regulators assigned to each split (*nreg*).

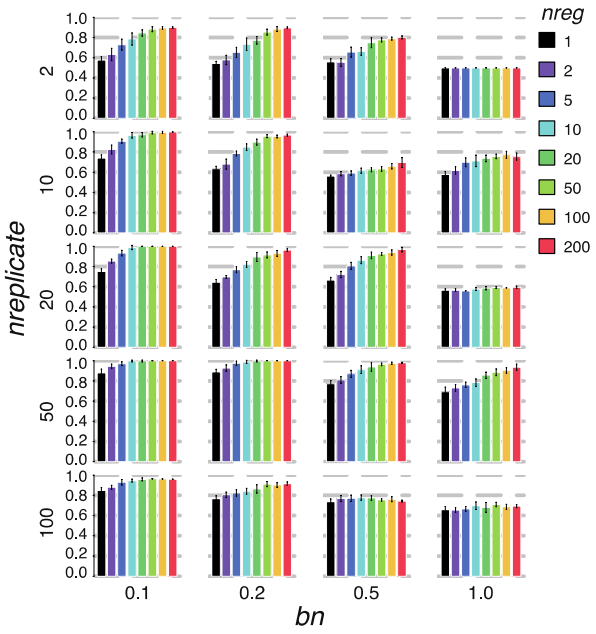

Fig. 2 AUC (TPR-FPR curves) of regulators assignment, rows *bn, c*olumns *nreplicate*. For each subfigure, *y*-axis AUC, *x*-axis, colors *nreg.*

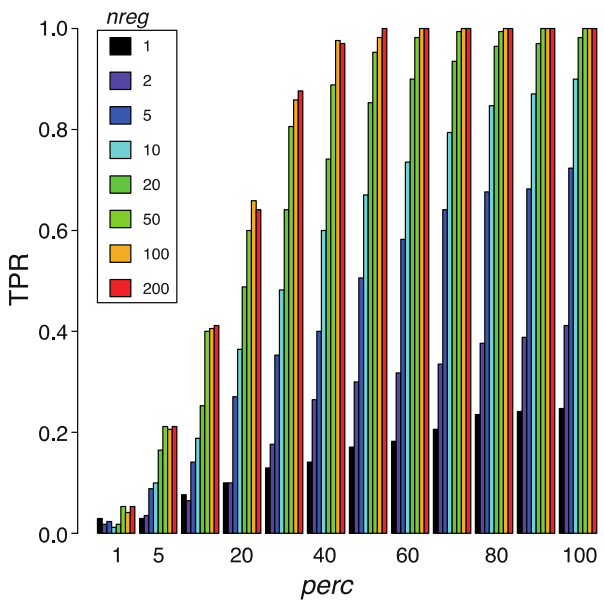

Fig. 3 TPR dependency of regulators assignment given *bn* = 0.1, *nreplicate* = 2. *x*-axis *perc*, *y*-axis *TPR*, colors *nreg*. For the complete figure see ESI,† Data S4, Figure S4-4.

Analysis by linear regression shows that all 3 variables significantly correlate with AUC (*p*-value 0.0002, 2e-16, 0.001, respectively), suggesting that each parameter contributes to the overall performance. In particular, we observe, first, that the positive correlation between *nreg* and AUC can be perceived in virtually all datasets at relatively low noise levels (*bn* = 0.1, 0.5, Fig. 2, left side columns). For very few instances with higher noise levels, such as *nreplicate* = 100, *bn* = 0.5 and 1.0, this trend nearly vanishes (for *nreplicate* = 2, *bn* = 1.0 the clustering failed to generate any experiment (sample) partition, AUCs are hence constantly 0.5). Second (Fig. 2, right side columns), the degeneration of AUCs with high noise levels is also obvious and expected (with the exception *nreplicate* = 10 groups, *bn* = 1.0, to the failure in generating experiment (sample) partitions under extremely high noise). Third (Fig. 2, rows from top to bottom), larger sample size is generally associated with better AUC, in spite of a few exceptions (*nreplicate* = 10, *bn* = 0.5 is lower than *nreplicate* = 2, possibly resulting from the instability of the regulators assignment algorithm as we observed before).

In this context, it is worthwhile warning the users on the peculiar meaning of the "random regulator list" output by LemonTree alongside the "assigned regulator list" for each module. For the generation of the former, assigned regulators are sampled at random from all regulators' scores, but these scores are not permutated across regulators. As a result, randomly assigned regulators also demonstrate reasonable performances (most AUCs > 0.5, ESI,† Data S4, Fig. S4-1), although generally lower than the true counterparts (e.g. in *nreplicate* = 2 or 10 groups). This peculiar definition of randomness must be known as it produces unexpected results when comparing such "random" lists with the non-random ones using ranking-aware analyses such as gene set enrichment analysis (GSEA,[19]), as the regulators still preserve the ranking information of assigned regulators as in the non-random lists. More details on TPR–FPR dependency using ROC curves, as well as ROC profiling of TPR-FDR can be found in ESI,† Data S4, Section 2, Figures S4-2,3.

The ROC curves (ESI,† Data S4, Figures S4-2, S4-3) were obtained by varying the parameter *perc* to get a series of TPRs/FPRs/FDRs at each threshold and finally plotting one variable (e.g. TPR) against the other (e.g. FPR), these results therefore do not directly reflect the impact of *perc*. Hence, we characterized the dependency of TPR (Fig. 3) on *perc*. The general TPR-*perc* trend shows that TPR rises with *nreg* as well as *perc*, and decreases with *bn* (most notably in *nreplicate* = 2 and 100 datasets). In *nreplicate* = 10, *bn* = 1.0, and *nreplicate* = 50, *bn* = 0.5 datasets, TPRs are abnormally higher than their less noisy counterparts, which is possibly due to the algorithmic flaw in regulators assignment procedure, where more distinct expressions between a node (i.e. in less noisy cases) may produce worse assignments (see Subsection **Improvement of regulators assignment for candidates with opposite expressions in a split**). These exceptions represent a more theoretical than practical issue as such unusually high noises (*bn* ≥ 0.5) should rarely happen. It is noteworthy that, although LemonTree documentation sets as default *perc* = 1%, our results show that this can lead to low sensitivity, as TPR tends to be < 0.1 even for large datasets (*nreplicate* = 20, 50,

100). As per our synthetic data analysis, *perc* should be set at least to 30%, in order to achieve a decent TPR (> 0.6).

We hypothesized that high FDRs could be overcome by a stringent *perc*. To test this, we profiled the dependency of FDR on *perc* as well (Fig. 4). Nevertheless, in accordance to the TPR-FDR ROC curves, *nreplicate* = 2 datasets are constantly plagued by high FDRs, even when *perc* = 1%. Therefore, we conclude that FDR is an important factor and should be controlled with high priority when applying LemonTree to datasets with limited sample size. In comparison, for datasets with modest sample size and little noise (*nreplicate* = 10 ~ 50, *bn* ≤ 0.2), FDR is no longer an issue even when *perc* = 100% (the full "assigned regulator list"): for example, when *nreplicate* = 10, *bn* ≤ 0.2, we recommend *perc* = 20% for a TPR ≈ 0.6 and FDR < 0.2, a good balance between sensitivity and FDR. Lastly, we observed generally higher FDRs in large datasets (*nreplicate* = 100); FDRs increase monotonically with the noise level. As a consequence, too many replicates (*nreplicate* ≥ 100) should also be dealt with caution in case of large clinical trials that can reach 100 replicates per treatment/condition.

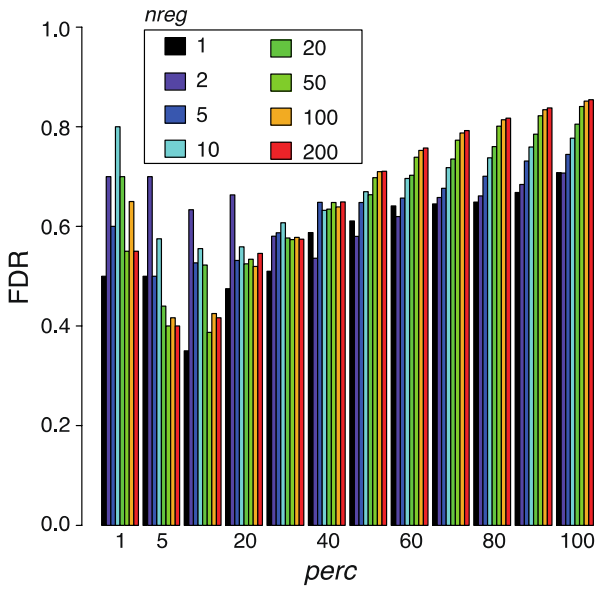

Fig. 4. FDR dependency of regulators assignment given *bn* = 0.1, *nreplicate* = 2. *x*-axis *perc*, *y*-axis *TPR*, colors *nreg*. For the complete figure see ESI,† Data S4 Text, Figure S4-5.

Regulators assignment using real data

LemonTree is designed for producing a set of assigned regulators in descending order of likelihood. Besides accuracy, the quest for a stable final solution requires the identification of: i) the *optimal* number of most likely regulators to be output (*perc*); ii) the *optimal* settings for *reg* and *nreg*. This section responds to such queries. By "optimal" we mean that the corresponding setting grants considerations for three variables equally important in the reliability of the solutions: 1) the stability of the assigned regulators, 2) the sensitivity and 3) the specificity of regulators assignment.

As shown in Fig. 5, stability is strongly affected by the number of candidate regulators. In case of short candidate regulator lists (e.g. *reg* = 10) substantial variability in the stability across modules can be perceived (column 1, rows 1-3) when we select the top regulators (*perc* within the range 1% ~ 10%) from the full output list. Therefore, LemonTree's default *perc* = 1%, showing the most extreme variability, should be revised in light of these results.

We also observe that at a fixed *perc* stability tends to decrease as the candidate regulator list gets longer (e.g. *perc* = 100%). This can be explained by the fact that the search space expands according with the growing number of candidates, hence the probability to have overlapping solutions by chance drops. As the number of candidate regulators increases, we observe that differences in stability between different *nreg* become more obvious. This trend can be further highlighted and summarized by introducing a new parameter *preg* (= *nreg*/*reg*). We noticed that when *nreg* < *reg* (e.g., *nreg* = 10, 50 given *reg* = 500), LemonTree can only output at most *nreg* of the assigned regulators (sampled with replacement). Thus, the lower *preg*, the lower the stability. This explains the gradually lower stabilities when *nreg* changes from 500 to 5 in *reg* = 500, *perc* = 50% (Fig. 5, central rows). In the performance tests using synthetic data, we have shown the best AUCs of *nreg* = 100 and 200 (i.e. *preg* = 100/117 and 200/117, both approximately ≥ 1) in comparison with smaller *nreg* for most datasets, therefore it is advisable for *nreg* to be set at a value not lower than the number of candidate regulators *reg* in general. Only on very peculiar occasions, when controlling FDR is the priority, can we reduce *nreg* to be less than *reg,* at the sacrifice of sensitivity, inevitably.

All recommendations emerging from these analyses are summarized in Table 2 and applied to the RA GI microbiome dataset in the last subsection.

Improvement of regulators assignment for candidates with opposite expressions in a split

Ideally, candidate regulators with distinctly *opposite* expressions at a split (ESI,† Data S5, Section 1) exhibit expressions highly correlated to the genes in the corresponding module, and thus are expected to be true regulators. We observed that such candidates could be missed in the output. We found this to originate from a flaw in the bisection procedure (used for identifying the optimal β that maximizes the posterior probability of a regulator's expressions to predict the partition of conditions at a given node), causing these candidates to be excluded before the score evaluation step. We patched the program and validated our modifications. For a detailed theoretical analysis and code improvement see ESI,† Data S5, Sections 1-4.

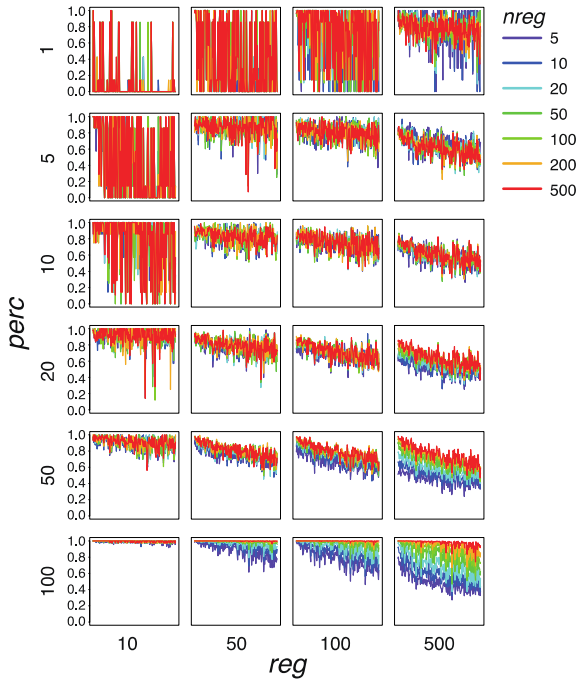

Fig. 5 Impact of the parameters *reg (x-axis), perc (y-axis) and nreg (color code)*, on the regulators stability (AOR).

Validation of the parameter recommendations using gut-intestinal metagenomic data

RA is an autoimmune disorder characterized by chronic inflammation, with an incidence (1% currently) growing worldwide[20], and no cure. RA, along with other autoimmune disorders, is known to be accompanied by *dysbiosis*[21, 22], i.e. the alteration of the composition of the gut-intestinal (GI) microbiome, an emerging player in the regulation of immune and metabolic functions. In particular, the possibility to manipulate the GI microbiome composition with modification of dietary habits and drug therapies[23, 24] (leading to the emerging field of nutraceutics) or via fecal microbial transplant (FMT).[25]

Very little is being done so far to evaluate the impact of therapies on the GI microbiome. This represents a lack that should be filled, with particular relevance in maladies accompanied by dysbiosis such as RA. Module networks, with the ability to infer cause-effect relationships can be used to explore the potential regulatory relationships that model the changes in composition of the GI microbiome from early untreated stages (NORA) to chronic treated phases (MTX and prednisone) aiming at the control of RA. In this context, natural/intuitive regulators of the process (early intestinal triggers of the disease, see also Material and Methods) are represented by the differentially abundant bacteria characterizing NORA with respect to the healthy baseline HLT.[17] Despite the uniqueness of this published dataset, there are two major limitations in these data. First, the definition of regulators associates the only available variables (diagnosis and associated clinical parameters and GI microbiome data) irrespective of other unknown parameters that may cause the differences in the GI composition. Diet is among such unknown and yet potentially relevant variables. Second, these do not represent longitudinal data, nevertheless the progression of the disease allows to assume that different early triggers in the GI data (from NORA patients) are potentially able to induce the different GI compositions observed at later stages (chronic patients) in the modules computed on MTX and prednisone patients' data.

Given the microbial abundance, GI microbiome regulatory module (Fig. 6A and Fig. 6B) was inferred with the parameters as described in ESI,† Data S3, Section 3. The algorithm produces only one tight cluster of genera from later stage treated patients (MTX and prednisone, the output module in Fig. 6B) and the associated set of regulators extracted from NORA patients (Fig. 6A).

The three regulating NORA genera (*Roseburia*, *Veillonella* and *Faecalibacterium*) have the ability to group patients displaying the clinico-pathological signs of NORA into two major patterns (bluish left-hand side, yellowish right-hand side, separated by a pink line, Fig. 6A). The information available in literature about these genera are relevant to gain more insight for these enterotypes. In particular, *Veillonella* and *Faecalibacterium* reductions are known dysbiotic characteristics of RA and Crohn's disease [25, 26] and *Roseburia*, among the better known and important butyrate producers, is fundamental in the control of inflammation by direct action on NF-κb.[27] The relative and coherent abundance of all three genera in the right-hand side subset of patients, and their deficiency in the left-hand one, allows to characterize the latter as patients missing important components to fight (chronic) inflammations, a hallmark of RA. Therefore, despite a common NORA diagnosis, patients on the left-hand side have, from the GI microbiome viewpoint, a stronger dysbiosis ($NORA_{dysbiotic}$) compared to the NORA samples on the right hand-side ($NORA_{eubiotic}$).

Table 2 Summary of LemonTree parameters' default values and our recommendations.

Parameter	Default/Exemplified	Suggestions
nc	30, 50	≥ 30
nreg/reg	NA	≥ 1
perc	1%	≥ 30%

The table compares the parameters' default settings or values used in LemonTree's publications with our recommendations. 'NA' denotes not documented.

Owing to the implementation of the module networks algorithm, such dual regulators' composition can be associated to, and further can be assumed to infer, a similarly shaped regulated output module with, on the left-hand side, low abundances (regulated by $NORA_{dysbiotic}$ genera) characteristic of MTX patients only (Fig. 6C), and on the right-hand side diverse and higher abundances, associated to prednisone and MTX patients.

To understand the relevance of this result, a digression on RA therapies is necessary. Prednisone is generally used at very low doses as a bridge drug, before or in support to DMARDs

(such as MTX) to control joints degeneration while limiting DMARDs' adverse effects.[28] Due to the well assessed advantages and recommendation to treat RA patients as early as possible to stop degeneracy irreversible progression, prednisone is recommended to patients with a less aggressive disease than the ones oriented to MTX.[29] It therefore comes with no surprise that $NORA_{dysbiotic}$ are associated to later MTX samples, and that prednisone patients are associated with $NORA_{eubiotic}$. I.e. the eu/dysbiotic landscape is predictive of the aggressiveness of the disease and of the associated therapy.

What remains to be explored and retains the most interesting finding is the group of MTX patients falling into the right-hand side of the module ($MTX_{NORAeubiotic}$ including samples MTX_0, MTX_1, MTX2, MTX_6). From our observations above, MTX patients should present with more aggressive signs of the disease, associated to an early dysbiotic landscape ($NORA_{dysbiotic}$). From the clinical point of view, $MTX_{NORAeubiotic}$ patients do not appear to differ from the other MTX patients (see Fisher's exact test, *t*-test as well as PCA analysis in ESI,† Data S3, Section 2). Yet, our results, upon microbiological analysis of the module output by LemonTree, suggest that such patients evolve towards different GI microbial compositions at later stages, with a visibly higher abundance (in comparison to prednisone patients) of *Parabacteroides*, including *P. distasonis* an opportunistic pathogen (Human Opportunistic Pathogens Library) and *P. goldsteinii* related to abdominal sepsis peritonitis.[30-32]

Globally we observe that dysbiotic patients at the time of early RA ($NORA_{dysbiotic}$) benefit from the treatment with MTX; similarly more eubiotic patients at the time of early RA ($NORA_{eubiotic}$) benefit from the treatment with prednisone; however, failing to assess the early (NORA) GI microbiome composition can lead to the inappropriate treatment of patients with MTX ($MTX_{NORAeubiotic}$), leading in turn, as shown in our results, to an exacerbation of dysbiosis. Based on what is known from the mechanisms at work in such a complex situation, we can only speculate that the reactive oxygen and nitrogen species (ROS, RNS) produced by the metabolism of MTX lead to adverse effects that represent a necessary price to pay in the case of a compromised microbiome.[33] However, MTX action in more eubiotic patients at the time of NORA is possibly and unnecessarily exacerbating the ROS and RNS production, contributing to the creation of an environment favourable to opportunistic bacteria and challenging the presence of obligate ones [34], limiting the endogenous ability of the patients to fight against inflammation. Although proper validation would require longitudinal data (from NORA to treatment on the same patient), these results offer preliminary indication with a clinical potential, inferred *in silico* by the powerful approach offered by module networks.

Fig. 6 GI microbiome regulatory module with genera in the rows (from phylum down to genus level for better annotation of the unclassified genera) and patients by therapy in the columns. The upper panel represents the heatmap of the assigned regulator microbes, the lower the heatmap of the regulated microbes. The red line segregates the patients into left/right cluster.

These findings are dramatically dependent on the careful parameters characterization we have described. In fact, the original default *perc* setting (1%) outputs an empty list (ESI,† Data S3, Fig. S3-2) missing the regulators (*Roseburia*, *Veillonella*, *Fecalibacterium*) herein identified, with the potential to improve the therapeutic management of RA. In particular, within the suggested range (*perc* ≥ 30% for an acceptable TPR), we show that *perc* = 70% leads to a stability ≥ 80% (ESI,† Data S3, Fig. S3-2), and that strikingly *Roseburia* appears as a constant top regulator in every instantiation, thus clearly identifying a highly probable regulator that would survive more stringent *perc* (10~20%) (ESI,† Data S3, Table S3-4).

Conclusions

In this study, we surveyed the parameter landscape of a freely available module networks implementation – LemonTree. In particular, we show its generally good performance for datasets that have a medium sized sample (*nreplicate* = 10 ~ 50) and acceptable noise range (*bn* ≤ 0.5). We inform users about the correct interpretation of the "random regulator list" as well as the imperative of properly controlling FDR in small datasets. Based on this, we provide users with suggestions on parameter settings for optimal performances (Table 2). Finally, we offer an algorithmic improvement to the regulators assignment procedure.

As we have shown in the benchmarking section, LemonTree's performances are indeed dependent on the data characteristics, and especially on the accompanying noise. The goal of this study is to provide users with full awareness on this dependency and parameters guidance under relevant experimental circumstance (i.e. a moderate noise level), where the results can be optimized by software's parameter settings. In particular, in our real-world GI-microbiome data application, we tested and verified the impact of *perc* on regulators' assignment stability. Based on this all, LemonTree's application to clinically relevant data offers the opportunity to speculate on therapeutic decisions based on informed GI microbiome analyses, demonstrating the efficacy and usability of our parameter recommendations and of module networks in general.

Acknowledgements

We thank LemonTree developers and anonymous reviewers for helpful communications and comments. We thank JE Dent for fruitful discussion on data presentation.

References

1 E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller and N. Friedman, *Nat Genet*, 2003, **34**, 166-176.
2 E. Segal, D. Koller, N. Friedman and T. Jaakkola, 2005.
3 G. Meng, A. Mosig and M. Vingron, *Bmc Bioinformatics*, 2010, **11**, 267.
4 C. J. Needham, J. R. Bradford, A. J. Bulpitt and D. R. Westhead, *PLoS computational biology*, 2007, **3**, 1409-1416.
5 Y. Lubling and E. Segal, Genomica, Segal Lab, Weizmann Institute, Rehovot, Israel, 2010.
6 E. Bonnet, L. Calzone and T. Michoel, *PLoS computational biology*, 2015, **11**, e1003983.
7 E. Bonnet, M. Tatari, A. Joshi, T. Michoel, K. Marchal, G. Berx and Y. Van de Peer, *Plos One*, 2010, **5**, e10162.
8 A. Joshi, R. De Smet, K. Marchal, Y. Van de Peer and T. Michoel, *Bioinformatics (Oxford, England)*, 2009, **25**, 490-496.
9 E. Bonnet, T. Michoel and Y. V. d. Peer, *Bioinformatics*, 2010, **26**, i638-i644.
10 V. Vermeirssen, A. Joshi, T. Michoel, E. Bonnet, T. Casneuf and Y. V. d. Peer, *Molecular bioSystems*, 2009, **5**, 1817-1830.
11 T. Michoel, S. Maere, E. Bonnet, A. Joshi, Y. Saeys, T. Van den Bulcke, K. Van Leemput, P. van Remortel, M. Kuiper, K. Marchal and Y. Van de Peer, *Bmc Bioinformatics*, 2007, **8 Suppl 2**, S5.
12 S. Roy, S. Lagree, Z. Hou, J. A. Thomson, R. Stewart and A. P. Gasch, *PLoS computational biology*, 2013, **9**, e1003252.
13 L. Rokach, *Artificial Intelligence Review*, 2009, **33**, 1-39.
14 T. V. d. Bulcke, K. V. Leemput, B. Naudts, P. v. Remortel, H. Ma, A. Verschoren, B. D. Moor and K. Marchal, *Bmc Bioinformatics*, 2006, **7**, 43.
15 J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz and T. R. Golub, *Nature*, 2005, **435**, 834-838.
16 N. Guelzim, S. Bottani, P. Bourgine and F. Képès, *Nat Genet*, 2002, **31**, 60-63.
17 J. U. Scher, A. Sczesnak, R. S. Longman, N. Segata, C. Ubeda, C. Bielski, T. Rostron, V. Cerundolo, E. G. Pamer, S. B. Abramson, C. Huttenhower and D. R. Littman, *eLife*, 2013, **2**, e01202.
18 V. Tremaroli and F. Backhed, *Nature*, 2012, **489**, 242-249.
19 A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov, *Proceedings of the National Academy of Sciences of the United States of America*, 2005, **102**, 15545 -15550.
20 Y. Alamanos and A. A. Drosos, *Autoimmun Rev*, 2005, **4**, 130-136.
21 J. Chen, K. Wright, J. M. Davis, P. Jeraldo, E. V. Marietta, J. Murray, H. Nelson, E. L. Matteson and V. Taneja, *Genome Med*, 2016, **8**.
22 G. B. Rogers, *Nat Med*, 2015, **21**, 839-841.
23 X. Zhou and C. Nardini, *BMC systems biology*, 2016, **10**, 107.
24 C. Nardini, V. Devescovi, Y. Liu, X. Zhou, Y. Lu and J. E. Dent, *Scientific reports*, 2016, **6**, 39043.
25 X. Zhang, D. Zhang, H. Jia, Q. Feng, D. Wang, D. Liang, X. Wu, J. Li, L. Tang, Y. Li, Z. Lan, B. Chen, Y. Li, H. Zhong, H.

Xie, Z. Jie, W. Chen, S. Tang, X. Xu, X. Wang, X. Cai, S. Liu, Y. Xia, J. Li, X. Qiao, J. Y. Al-Aama, H. Chen, L. Wang, Q.-J. Wu, F. Zhang, W. Zheng, Y. Li, M. Zhang, G. Luo, W. Xue, L. Xiao, J. Li, W. Chen, X. Xu, Y. Yin, H. Yang, J. Wang, K. Kristiansen, L. Liu, T. Li, Q. Huang, Y. Li and J. Wang, *Nat Med*, 2015, **21**, 895-905.

26 H. Sokol, B. Pigneur, L. Watterlot, O. Lakhdari, L. G. Bermúdez-Humarán, J.-J. Gratadoux, S. Blugeon, C. Bridonneau, J.-P. Furet, G. Corthier, C. Grangette, N. Vasquez, P. Pochart, G. Trugnan, G. Thomas, H. M. Blottière, J. Doré, P. Marteau, P. Seksik and P. Langella, *Proceedings of the National Academy of Sciences*, 2008, **105**, 16731-16736.

27 S. H. Duncan, G. L. Hold, A. Barcenilla, C. S. Stewart and H. J. Flint, *Int J Syst Evol Micr*, 2002, **52**, 1615-1620.

28 M. C. van der Goes, J. W. G. Jacobs and J. W. J. Bijlsma, *Current opinion in rheumatology*, 2016, **28**, 289-296.

29 M. D. Mjaavatten and V. P. Bykerk, *Best Practice Research Clinical Rheumatology*, 2013, **27**, 451-466.

30 F. M. Awadel-Kariem, P. Patel, J. Kapoor, J. S. Brazier and E. J. C. Goldstein, *Anaerobe*, 2010, **16**, 223-225.

31 *Journal*.

32 P. C. Y. Woo, S. K. P. Lau, J. L. L. Teng, H. Tse and K. Y. Yuen, *Clin Microbiol Infec*, 2008, **14**, 908-934.

33 P. Tieri, X. Zhou, L. Zhu and C. Nardini, *Systems biology*, 2014, **2**, 59.

34 S. E. Winter, C. A. Lopez and A. J. Bäumler, *EMBO reports*, 2013, **14**, 319-327.