Running title: GRNs via conditional inference forests

Practical aspects of gene regulatory inference via conditional inference forests from expression data

Kyrylo Bessonov[1], Kristel Van Steen [1]

Keywords: biological interactions, gene regulatory networks, conditional inference forests

Affiliation

[1]

Medical Genomics,
GIGA-R, Université de Liège,
Avenue de l'Hôpital, 1
4000, Sart-Tilman
Belgium

Corresponding author:

Kyrylo Bessonov kbessonov@ulg.ac.be

Tel. +32 48 863 8117

Abstract

Gene regulatory network (GRN) inference is an active area of research that facilitates understanding the complex interplays between biological molecules. We propose a novel framework to create such gene regulatory networks, based on Conditional Inference Forests (*CIF*s) as proposed by Strobl *et.al.* Our framework consists of using ensembles of Conditional Inference Trees (*CIT*s) and selecting an appropriate aggregation scheme for variant selection, prior to network construction. We show on synthetic microarray data that taking the original implementation of *CIF*s with conditional permutation scheme (CIF_{cond}) may lead to improved performance compared to Breiman's implementation of Random Forests (*RF*). Among all newly introduced *CIF*-based methods and 5 network scenario's obtained from the DREAM4 challenge, CIF_{cond} performed best. Networks derived from well-tuned *CIF*s, obtained by simply averaging *p*-values over tree ensembles (CIF_{mean}) are particularly attractive, since they combine adequate performance with computational efficiency. Moreover, thresholds for variable selection are based on significance levels for *p*-values and, hence, do not need to be tuned. From a practical point of view, our extensive simulations show the potential advantages of CIF_{mean} -based methods. Although more work is needed to improve on speed, especially when fully exploiting the advantages of conditional inference trees in the context of heterogeneous and correlated data, we have shown that *CIF* methodology can be flexibly inserted in a framework to infer biological interactions. Notably, we confirmed biologically relevant interaction between IL2RA and FOXP1, linked to the IL-2 signaling pathway and to type 1diabetes.

Introduction

Real-life biological systems display interactions and regulation schemes that are part of complex pathways or networks. Understanding these networks is important to unravel gene regulatory mechanisms or the genetic basis of complex disease traits. The availability of genome-wide transcriptome data offers opportunities and challenges for data analysts to extract gene regulation information directly from gene expression profiles: genes regulate each other’s expression and activity.

One of the challenges when dealing with data derived from high-throughput technologies (i.e., ‘omics’ data) involves the curse of dimensionality. This refers to the fact that number of variables *p* is usually much larger than the number of samples *n* for these data and hence model parameter estimation becomes unstable. Ignoring the *p*>>*n* issue and adhering to classical statistics, is bound to generate singularities in matrix algebra (e.g., singular matrices) [Johnstone and Titterington 2009]. The curse of dimensionality particularly applies to transcriptome data derived via RNA-*seq*, but also holds true for microarray-based data that typically considers between 10,000 and 57,000 transcripts, depending on the platform and organism [Hardiman 2004]. One way to circumvent this problem is to reduce the number of variables. This can be done by using prior biological knowledge leading to biologically motivated constraints, or via mathematical/statistical variable selection algorithms. Alternatively, novel representations of the data are looked for, such as principal components in a lower-dimensional linear space [Yao, et al. 2012] or kernels for non-linear data dimensionality reduction [Lin, et al. 2011].Graph structures are easy to interpret and naturally represent biological networks. These graph structures naturally give rise to biological networks [Zhu, et al. 2007]. These networks may refer to genes and gene products or to networks between macro- and/or micro-molecules, possibly integrating different data sources [Sahni, et al. 2015]. Although only approximate, biological networks are often assumed to be scale free. This implies that only a small number of nodes in the network are highly connected and that the majority of nodes are connected to only a few neighboring nodes. Usually, connected nodes in such networks are said to be “interacting”. However, this does not necessarily mean that the nodes (or the compounds they represent) are physically interacting. Note that several so-called physical interaction networks may miss true interactions as well and often contain non-functional interactions [Levy, et al. 2009]. Gene regulatory networks (GRNs) represent directed functional linkages existing between genes and regulatory elements most frequently associated with transcription factors [Davidson and Levin 2005].

In this work, the envisaged biological networks are functional GRNs for which “interactions” depict either direct or indirect regulatory relationships [Cho, et al. 2012]. The framework we develop relies both on tree-based variable selection and GRN inference. One of the advantages of trees and ensembles of trees [Huynh-Thu, et al. 2010] is their ability to effectively and rapidly dissect complex data spaces such as those generated by gene expression microarrays. Trees and random forest methodology belong to the class of recursive partitioning methods, that aim to recursively partition the space spanned by all input variables into partitions of observations with similar responses. The final partitions may be characterized by highly complex interactive patterns between input variables, although care has to be taken when interpreting interactions in the context of random forests [Boulesteix, et al. 2015]. For a general overview on classification and regression trees, we refer to [Loh 2011].

Tree-based models have a tendency to over-fit the data at hand and, hence, to underestimate classification errors. Several measures can be taken to overcome these issues, including the building of unpruned trees on multiple bootstrap samples as implemented in Breiman Random Forests (*RF*) [Breiman 2001] and the separation of variable selection and node splitting steps [Strobl, et al. 2007]. Such a separation is implemented in Conditional Inference Trees (*CIT*) and Conditional Inference Forests (*CIF*s) [Strobl, et al. 2008; Strobl, et al. 2007]. At the heart of *CIT* and *CIF* lies an unbiased tree algorithm that do not artificially favor splits in variables with many categories or continuous variables [Strobl, et al. 2009]. *CIF*s present several advantages over classical *RF*s including separation of node selection and splitting steps to overcome tree-based variable selection bias [Hothorn, et al. 2006], resampling with replacement to handle ensemble variable selection bias introduced by bootstrap sampling [Strobl, et al. 2007], a conditional permutation scheme to deal with correlated input features [Strobl, et al. 2008], and the possibility of natural threshold selection for variable importance measures (*VIMs*), as we will show later. Hence, a *CIF*-based methodology theoretically encompasses categorical and continuous input variables that are possibly inter-related and measured on different scales, hereby paving the way for combined analysis of multiple data sources. For these reasons, and having integrative analyses of heterogeneous and inter-connected ‘omics’ data in mind, we chose *CIT*s and *CIF*s as the basis of our novel network construction and inference methodology, despite the fact that random forests rather than *CIT*s or *CIF*s are widely applied in bioinformatics contexts [Boulesteix, et al. 2012b]. In the belief that computational efficiency can be reached by optimizing the program code and *CIT*/*CIF* underlying algorithms, we focus on investigating the impact of parameter choices in *CIT* and *CIF* (e.g., related to multiple testing correction and the number of randomly selected variables at each tree node) on the performance of proposed gene regulatory network construction methods, in synthetic and real-life data settings.

The source code: The source code developed in the context of this manuscript is available at https://bitbucket.org/kbessonov/cifmean and www.statgen.ulg.ac.be

Methods

Data sources

We obtained publicly available gene expression data from the DREAM 2, 4 and 5 challenges [Marbach, et al. 2012; Marbach, et al. 2010; Marbach, et al. 2009] and the GEO public repository (GEO #: GSE43488).

In particular, we used gene expression data on 3456 *E.coli* genes from DREAM2, containing 320 transcription factors (TF), for 300 subjects [Essaghir, et al. 2010]. The 320 TFs were considered as input variables to our proposed strategies (Figure 1). As the gold standard (GS) network we took evidences from RegulonDB [Salgado, et al. 2013] of experimentally verified regulator - target gene relationships.

The DREAM4 In Silico Network Challenge data only contained synthetic microarray expression data derived from 5 GS networks, each with 100 nodes [Marbach, et al. 2010; Marbach, et al. 2009; Prill, et al. 2010]. Each data set contained 100 genes collected on 100 samples. Since no list of potential regulators existed, all 100 genes were considered as possible input variables.

The DREAM5 Network Inference Challenge data consisted of three GS networks 1-3, with respectively 1643, 4511 and 5950 genes. The GS network 1 data contained synthetic (simulated) gene expression data represented by 1643 genes and 195 regulators for 805 samples. The GS network 2 real-life *E.coli* expression data was characterized by 4511 genes containing 334 TFs collected on 805 subjects. The GS network 3 also involved real-life data on an organism, this time *S.cerevisiae.* The corresponding gene expression data included 5950 genes containing 333 regulators and was collected on 536 samples. For each scenario, the entire set of regulators was used as starting set for variable selection in our methodology.

As a case study, we took human microarray expression data from a type 1 diabetes (T1D) study in children [Kallionpaa, et al. 2014], obtained via the public GEO database (GEO #: GSE43488). As gold standard we considered the verified set of transcription factor–target gene sets from [Essaghir, et al. 2010] that used a variety of sources, including the Transcriptional Regulatory Element Database (TRED) [Jiang, et al. 2007], Pazar [Portales-Casamar, et al. 2009], PubMed, and the Transcription Regulatory Regions Database (TRRD) [Kolchanov, et al. 2002], among others. The resulting unique list of gene-gene pairs was composed of 1617 genes (245 TFs and 1372 target genes). These 1617 genes, evaluated on 121 samples, served as input to the analytic tools considered in this manuscript. A summary of the available data is given in Table 1.

CIT/CIF-based network inference methodologies

A schematic representation of our proposed GRN framework is given in Figure 1. In particular, for a given ‘omics’ data set, with molecular information that can be mapped to a gene, for instance transcriptome data, and assuming a one-to-one mapping of transcripts to genes, each transcript (gene) is subsequently taken as output (response) and the remaining transcripts (genes) are taken as input (predictor variables). For each response, a *CIF/CIT* is constructed and a variable importance measure (*VIM*) is defined (see below). These measures per gene are either based on a single *CIT* or are aggregated over several *CIT*s in gene-based *CIF*s, depending on the view taken to construct a network from trees. In general, a (statistically) "significant" *VIM* for gene *X* in predicting gene *Y* will lead to a connection between *X* and *Y* in the network. Because of the direction of prediction, the connection is presented as a directed edge, naturally giving rise to a directed network (i.e., GRN). The so-called predicted network is compared to a gold standard (when available), using network prediction performance criteria as suggested by [Marbach, et al. 2012; Prill, et al. 2010]: 1) the area under the receiver operating characteristic curve (*AUROC*), 2) the area under the precision-recall curve (*AUPR*), and 3) the *DREAM* challenge specific score. The ROC curve plots the sensitivity (i.e., true positive rate) versus 1 minus specificity (i.e., 1 minus the true negative rate) and is well-known in statistics. Precision-Recall curves or PR curves are often used in Information Retrieval and offer an alternative to ROC curves for skewed class distributions. An algorithm may be a good performer based on ROC but not based on PR. Whereas recall is defined as the true positive rate, precision is defined as the fraction of examples classified as positive that are truly positive. When the number of unconnected nodes exceeds the number of connected nodes in the GS networks, as is the case with GRNs, more information about comparative performance of methods can be retrieved from precision-recall curves [Davis and Goadrich 2006]. For more details about ROC-PR comparisons, we refer to [Davis and Goadrich 2006]. The overall score summarizes performance over several network scenarios and is defined as in [Marbach, et al. 2012] as the mean of the (minus log10-transformed) network specific *p*-values p_{PR} and p_{ROC}. The PR and ROC *p*-values are derived from the original *AUPR* and *AUROC* values by comparison of obtained areas with those obtained from a simulated null distribution based on 25,000 random networks [Marbach, et al. 2012].

In what follows, we briefly describe the network inference schemes considered in this work. Each of these schemes involved particular choices of *VIM*s and hence different gene-gene network building strategies:

CIT: Here, the global null hypothesis of independence between any of the predictors and the response under consideration is tested by means of the conditional distribution of linear statistics in the permutation test framework of [Strasser and Weber 1999]. When this hypothesis cannot be rejected, the procedure stops. Otherwise, the predictor with the strongest association to the response is selected. We define the node's variable importance measure as its measure of association with the response (i.e., the *p*-value of the corresponding association test) and denote it as VIM_{node}. When the same variable refers to multiple times nodes in the same tree, the measure of association corresponding to the node that exhibits the largest sample size is considered. Next, the most optimal split for that node is sought (i.e., the split that maximizes a split statistic). It is based on standardized linear statistics as before. For more details, we refer to [Hothorn, et al. 2006].

CIF* and *CIF_{cond}: In this work, both GRN inference schemes build an ensemble of conditional inference trees via the R *party* package version 1.0-11 [Hothorn, et al. 2014; Hothorn, et al. 2006; Strobl, et al. 2008]. The *cforest_control()* function therein defines parameters that control the tree building. Unless stated otherwise, we passed the following parameters described in [Hothorn, et al. 2014] to *cforest_control()*: *teststat="quad", testtype="Univariate", fraction=0.632, replace=F, mincriterion=0.95, minsplit=20, ntree=1000, mtry=k/3*. Note that *teststat="quad", testtype="Univariate",* and *replace=F* correspond to the recommendations given in [Strobl, et al. 2007], so as to construct unbiased random forest. The *mtry* parameter (i.e., the number of variables randomly selected at each node) was set to *k/3*, with *k* the total number of possible predictors in the data as recommended by [Boulesteix, et al. 2012a] . While the *cforest()* function creates ensembles of trees from a training section of the input data, the function *varimp()* uses the out-of-bag (*OBB*) samples to calculate the importance of each predictor variable with respect to target response.. In particular, for each gene predictor / gene response pair, the *varimp()* function outputs the mean decrease in accuracy (*%IncMSE*), indicating how much the mean square error (MSE) increases after permutation of the *OOB* samples averaged over all trees of the forest. Thus, large values of *%IncMSE* are suggestive of a gene pair's importance. Because for forests, a node's variable importance is aggregated over several trees, we denote it by *VIM*VIM_{global}. In practice, its calculation was made by the *varimp()* function with parameters *nperm=100* and *OOB=T*. Hence, we used a total of 100 data permutations and OOB samples in the testing phase. In the case of CIF_{cond} the *conditional* parameter in the *varimp()* function was set to *true* (i.e., variable importance was assessed via the conditional importance measure of [Strobl, et al. 2009]), while in *CIF* it was set to *false*.

CIF_{mean}: In contrast to *CIF* and CIF_{cond}, we passed the following tree growth parameters to the function *ctree_control(): teststat="quad", testtype="Univariate", fraction=0.632, replace=F, mincriterion=0.95, minsplit=20, ntree=1000, mtry=k/3*. Because in CIF_{mean} variable importance is assessed within a statistical framework, based on formal testing and *p*-values, the obtained *p*-values were compared to a significance level of 0.05 (i.e., *mincriterion=0.95*). In addition, a minimum number of 20 individuals were required in a node before it was considered for node splitting (*minsplit=20*). A node's variable importance VIM_{global} was aggregated over several trees according to the formula in Eq. 1, with $n(X_j)$ - the number of trees grown per response gene that contain the variable X_j as a node and p_{Xjt} the *p*-value related to the association test between predictor gene *j* and response gene *i* in tree *t* of the ensemble (VIM_{node}). As before, when gene *j* (X_j) occurs twice in the same tree, only the *p*-value corresponding to the largest sample node is considered. We use CIF_{mean} *p-value* to refer to network inference strategies in which VIM_{global} is calculated using *p*-values aggregated via the equation 1. In case *testtype="Teststatistic"* in the *ctree_control()* function above, not *p*-values but raw test-statistics are used to aggregate over trees. We refer to this strategy as CIF_{mean} *test-statistic*.

$$VIM_{global} = a_{ij} = \frac{\sum_t^T p_{Xj}^t}{n(X_j)} \text{ (Eq. 1)}$$

Breiman *RF*: We implemented classic random forest (building 1000 trees) with the *randomForest* library (version 4.6-7) in R [Breiman 2001; Liaw and Wiener 2002] and the default options with *mtry=k/3*. Similar to the *CIF* and CIF_{cond} methodologies described above, VIM_{global} importance measures were permutation based and reflect the mean decrease in accuracy (*%IncMSE*) before and after permutation of *OOB* samples.

The *OOB* samples were derived based on sampling with replacement (bootstrapping) equivalent to the *replace=T* in the *CIF*. They were computed via the function *importance(...)[,"%IncMSE"].*

The conditional inference framework and multiple-testing: Previously, we indicated that we based the stopping criterion during node selection in *CIT* or *CIF* on univariate (multiple testing uncorrected) *p*-values as invoked by *testtype="Univariate"* in *ctree_control()*. However, it is also possible to use a stopping rule based on test statistics rather than *p*-values. In comparison to the second, the first does not make assumptions about the nature of large-sample distributions. Currently, in the software, it is only possible to explicitly account for multiplicity in the node selection, when using a stopping rule based on *p*-values (*testtype="Univariate"*), either by relying on Bonferroni (*testtype="Bonferroni"*) or Monte Carlo (*testtype="MonteCarlo"*) strategies. In practice, with Bonferroni correction, a node's variable importance measure VIM_{node} is calculated using the formula $1\text{-}(1\text{-}p_{raw})^k$ where *k* is the total number of input predictor variables minus 1 [Hothorn, et al. 2014]. The Monte Carlo multiple-testing corrected *p*-values attached to a node (VIM_{node}) are based on permutations and adopts a *min-p* approach [Hothorn, et al. 2014].

In the next section, we report results of extensive simulation studies, using the aforementioned network inference schemes and assess their robustness to altered parameter choices. In addition, we explore their utility on real-life data applications and formulate recommendations of our proposed GRN framework in data integrative contexts.

Results

Evaluation of CIT/CIF-based GRN inference with DREAM4 data

Based on the DREAM overall score criterion (see Methods section), the best performers were CIF_{cond} (34.24) and *CIF* (33.92), followed by *RF* (33.50) and CIF_{mean} based on aggregating *test-statistics* rather than *p*-values (27.39) (Supplementary Figure S1). Amongst the CIF_{mean} methodologies based on *p*-value aggregation, the best performers were GRN methodologies that utilized mulitple testing (MT) correction. The Monte Carlo based MT correction was the most effective (23.75), closely followed by Bonferroni

(23.61). The prediction performance of GRNs derived from a single tree (*CIT p-value (Uncorrected)*), compared to CIF_{mean} *p-value (Monte Carlo)* was 1.8x lower: 13.17 compared to 23.75. The *AUROC* and *AUPR* in the 5 networks separately showed quite diverse performance trends amongst the considered GRN inference methods, as can be observed from Figure 2. However, CIF_{cond} is a quite computationally intensive strategy (Table 2). The computations shown in Table 2 were run on a single core of an Intel L5420 processor clocked at 2.50 Ghz. For hundred nodes and a sample size of 100 subjects, Breiman's *RF* implementation was the fastest data partitioning method, closely followed by CIF_{mean} *p-value (Bonferroni)*. For this reason, and because it is an easy-to-implement strategy giving rise to a statistically grounded threshold for variable importance, we will focus on CIF_{mean} in the remainder of this manuscript (DREAM2, DREAM5 and real-life data) and will investigate the scenario's under which the performance of CIF_{mean} can be optimized.

The parameter *mtry* can have a large impact on GRN inference performance, as can be seen from Figure 3 for CIF_{mean}. The highest DREAM4 overall scores were obtained for *mtry=k*, hence using all possible input predictors, with the exception of CIF_{mean} *p-value (Bonferroni)*. There *mtry*=5 seemed to be a reasonable choice.

Evaluation of CIF_{mean}-based GRN inference with DREAM2 data

The top performer based on *AUROC* and *AUPR* was *RF* followed by CIF_{mean} *test-stastistic (Uncorrected)* and CIF_{mean} *p-value (Monte Carlo)* (Figure S2). The Monte Carlo multiple-testing correction provided the best performance amongst the CIF_{mean} *p-value* methods (Supplementary Figure S2). Note that since only a single data scenario was available for DREAM2 it was not possible to compute a DREAM global score but instead, we considered the sum of *AUROC* and *AUPR*. Contrary to DREAM4 results, the optimal *mtry* parameter based on *AUPR* and *AUROC* across all CIF_{mean} methods with the exception of Bonferroni was at default value of *k/3*. In case of CIF_{mean} *p-value (Bonferroni)*, the highest performance was reached at the maximal possible value for *mtry* (Figure 4: *mtry=k*).

Application of *CIF*s to DREAM5 data

Using the DREAM5 overall score developed in [Marbach, et al. 2012], the best performers were *RF* (63.28), followed by CIF_{mean} *p-value (Monte Carlo)* (54.02) and CIF_{mean} *p-value (Uncorrected)* (53.59). Very low performance was demonstrated by CIF_{mean} *p-value (Bonferroni)* showing ~8.5x performance drop compared to CIF_{mean} *p-value (Monte Carlo)*: 6.32 compared to 54.02 (Figure S3 and Figure 5). The default *mtry* value of *k/3* provided the optimal performance for CIF_{mean} uncorrected for multiple testing and CIF_{mean} with Monte Carlo based multiple testing corrected *p*-values. For CIF_{mean} with Bonferroni corrections and for CIF_{mean} *test-statistic (Uncorrected)*, the maximal value of *mtry* provided the highest performance (Figure 6).

Case Study: T1D data

The best CIF_{mean} -based GRN performer, at a default of *mtry=k/3,* was CIF_{mean} *p-value (Monte Carlo)* followed by CIF_{mean} *p-value (Bonferroni)* (Supplementary Figure S4). The impact of the *mtry* parameter across CIF_{mean} methods greatly varied (Figure 7a). In case of the CIF_{mean} *p-value* method with Monte Carlo multiple testing correction, the highest performance was achieved at the default *mtry* setting of *k/3*, although the performance is rather stable across the considered values for *mtry*. With Bonferroni correction, the maximum value of *mtry* at *k* gave the highest performance benefits. The CIF_{mean} *p-value (Uncorrected)* method showed the lowest performance changes with varying *mtry* parameter values (Figure 7a), and therefore appears to be the most stable approach among all CIF_{mean} approaches. The default *mtry* value of *k/3* was clearly suboptimal for CIF_{mean} *test-statistic (Uncorrected)* and CIF_{mean} *p-value (Bonferroni),* depending on whether *AUROC* and *AUPR* measures of performance were considered (Figure 7b).

From a practical point of view, considering a threshold *p*-value of 0.01, the GRN inferred with the best performer CIF_{mean} *p-value (Monte Carlo)* with *mtry=k/3*, highlighted a total of 89 interactions. Amongst them is a highly significant pair involving forkhead box P1 (*FOXP1*) and the IL-2 receptor-α (*IL2RA*), with corresponding *p*-value based global variable importance measure of 0.0057 (Table 3). Both *IL2RA* and *FOXP1* are well known T1D markers linked to the IL-2 signaling pathway and immune regulation [Hulme, et al. 2012]. Table 3 lists other significant pairs linked to *IL2RA*.

Discussion

Networks come in different flavors, depending on their aim and the biological entities that serve as input during their construction. Examples of networks include gene regulatory networks (GRN) [Baitaluk, et al. 2012], co-expression networks [Horvath and Dong 2008; Langfelder and Horvath 2008; Wang and Huang 2014], differential networks [Ideker and Krogan 2012], metabolic networks [Maarleveld, et al. 2013]. All the networks considered in this manuscript were directed and inferred via machine learning based methodologies inspired by forests. Genes were taken as nodes and "variable importance measures" were taken as weights to edges. The measures were derived from conditional inference trees (*CIT*s) or conditional inference forests (*CIF*s). The reason for relying on conditional inference trees rather than classic regression trees was that we were ultimately interested in developing a network inference method that enables the integration of different data types (for instance, methylome, genome and transcriptome data). These data types generate measurements on differential scales, requiring re-scaling in order to avoid biased selection of features. Specifically, Breiman's Random Forests [Breiman 2001] are known to be biased towards features with larger number of possible splits [Strobl, et al. 2007]. In addition, correlations between features are frequent in biological data (e.g., co-expression networks rely on "correlations" between gene expressions). Rather than reducing the data to obtain independent features (e.g. via components theory which would complicate node definition and interpretation), a method that can directly deal with correlated features is highly desirable. Our results showed that the conditional inference forests (*CIF*) framework can outperform classic Random Forest, especially when features are correlated or are of different measurement types as was demonstrated in DREAM4 data (Figure 2 and Figure S1).

In particular, CIF_{cond} applied to relatively small data from the DREAM4 challenge (100 nodes and sample size of 100), outperformed all other considered methods based on *AUPROC* and *AUPR* performance measures, including *RF*. The added value of CIF_{cond} to *RF* seems to be rather small at first sight, but given its theoretical optimality in the presence of correlated data (as is the case here: multiple genes are co-expressed), we would generally favor CIF_{cond} over *RF*. Note that only weak correlation patterns existed between gene expressions in DREAM4 data. Averaged over all networks, only 2.20±0.91% of gene pairs showed a correlation > 0.3 (Figure S5). Interestingly, for network 4 the *AUROC* of *CIF$_{cond}$* was largely suboptimal to *RF*, whereas for *AUPR*, the *CIF$_{cond}$* slightly outperformed *RF* (Figure 2). Clearly, single tree-based techniques are not to be recommended for GRN inference purposes (Figure 2).

Interestingly, having a closer look at DREAM4 scenarios and Figure 2 and *AUPR*, *CIF$_{mean}$* with a stopping-rule based on test statistics rather than *p*-values outperformed all other *CIF$_{mean}$* methodologies. This may be due to the fact that *CIF$_{mean}$ test-statistic (Uncorrected)* does not make any assumptions about the shape or nature of the test statistic's distribution. Hence, it would be interesting to investigate in more detail the relation between GS network properties, the nature of the input variables and the performance of *CIF$_{mean}$ test-statistic (Uncorrected)*, possibly combined with a *maxT* [Westfall 1993] approach to derive multiple testing corrected *p*-values. The same observation was made for DREAM2 data (Figure 4 – varying *mtry* values).

Among the *p*-value based *CIF$_{mean}$* methodologies, *CIF$_{mean}$ p-value (Monte Carlo)* was the best performer for DREAM4 (Figure S1) and DREAM2 (Figure S2). For DREAM5 data scenarios, which are scenario's with the largest numbers of genes among all considered synthetic networks in this manuscript, *CIF$_{mean}$ p-value (Monte Carlo)* did not only outperform all *p*-value based *CIF$_{mean}$* methodologies, but also *CIF$_{mean}$* based on test statistics (Figure S3). In DREAM5, *CIF$_{mean}$ p-value (Monte Carlo)* was closely followed by *CIF$_{mean}$ p-value (Uncorrected)* (Figures 5 and S3). All these results seem to indicate the added value of adjusting for multiplicity during node selection, despite it being more computationally intensive (Table 2).

The most optimal *mtry* value highly depended on the data scenario, respectively DREAM2, DREAM4 and DREAM5 (Figures 3, 4 and 6). For *CIF$_{mean}$ p-value (Uncorrected)* methodology, the most optimal values were respectively *k* and *k*/3. For *CIF$_{mean}$ p-value (Bonferroni)* they were 5 and *k* with *k/3* being a reasonable alternative. For *CIF$_{mean}$ p-value (Monte Carlo)* the most desirable *mtry* values were *k* and *k/3* and for *CIF$_{mean}$ test-statistic (Uncorrected)* they were *k/3* and *k*. Notably, only with DREAM5 data the number of samples largely exceeds the number of the input variables (i.e., TFs) to be considered for gene network inference. Hence, DREAM5 more closely resembles a classic regression context, compared to the other data scenarios, for which it has been shown that there is little improvement by using unpruned bagging strategies (i.e., *mtry*=*k*). The higher the discrepancy between the number of samples compared to the number of input variables (*p*>>*n*; as is the case for most real-life omics data with human samples), the more we expect *mtry*=*k* to do well, as was observed for T1D data (Figure 7). For practical reasons, our *CIF*-based GRN inference framework takes *mtry*=*k/3* as a default.

From a theoretical point of view and on small data sets CIF_{cond} is to be preferred (Figure 2 and S1). From a practical point of view, more work is needed to use CIF_{cond} principles for GRN inference purposes. Based on the DREAM4 data represented by networks composed of (only) 100 nodes, the computation time of CIF_{mean} *p-value (Uncorrected)* was 12.35 minutes versus 0.79 minutes for *RF* (Table 2). Analyzing 4511 nodes of the DREAM5 network 2 took CIF_{mean} *p-value (Uncorrected)* 3232 minutes versus 6054 minutes for *RF*. Hence, it seems that the larger the data, with the same *mtry* parameters, the larger the computation time advantage of CIF_{mean} over *RF* may be. Clearly, as CIF_{cond} already took approximately 2 hours analyzing a 100 node network versus 12.35 minutes for CIF_{mean} *p-value (Uncorrected),* it is infeasible to use it on large data sets at the moment. Modifying CIF_{cond} to reduce computation time is the subject of future projects.

GRN inference in eukaryotic expression data is complex [Michoel, et al. 2009]. Therefore, the absolute drop in performance of CIF_{mean} *p-value (Monte Carlo)* (Figure 7), compared to for instance DREAM4 data (Figure 2), on type 1 diabetes (T1D) data is not surprising. Low correlation between gene expression levels is possibly due to transcription factor regulation acting on the protein rather than on the transcript level itself via post-translational modifications, unknown latent variables, genes exhibiting functional overlaps, several levels of regulation that are not caused by transcription factor binding [Hu, et al. 2007], epigenetic components and others. This may result in gold standard networks with heavy reliance on protein-protein binding but poor expression level changes [Marbach, et al. 2012]. Nevertheless, CIF_{mean} identified highly relevant T1D genes connected to IL2RA, a well-known marker of T1D. This suggests the potentials of

CIF_{mean} – based GRN inference in unraveling biologically relevant mechanisms. Among the genes listed in Table 3, is *STAT1*, a member of the STAT protein family, which is critical in IL2 signaling and regulation of T cell activity. Perturbations in IL2 signaling pathway were found to be closely associated to onset of T1D, highlighting the importance of the immune system and cytokine signaling components [Hulme, et al. 2012]. The FOX family of proteins and BCL3 highlight the immune system involvement in T1D. The SMAD family of proteins is also key to T1D, as they are associated with TGFβ and BMP pathways. Mutations in *SMAD* genes are strongly associated with diabetes, as previously reported in [McKnight, et al. 2009]. Note that these results were obtained by taking a *p*-value threshold of 0.01. The optimal threshold in the corresponding ROC curve (i.e., the point closest to the top-left part of the plot) was 0.0038 with respective specificity and 1-sensitivity confidence intervals of 0.55-0.56 and 0.45-0.52 (based on 2000 bootstrap samples).

Finally, *CIF*'s separate node selection and splitting association steps coupled to general association measure based on framework developed by Strasser and Weber [Strasser and Weber 1999] offer opportunities to handle different input data types, for instance RNA-*seq* and microarray expression data. Performance of *CIF* based methodologies was not yet tested on RNA-*seq* data, which is often characterized by small sample sizes. Since RNA-*seq* transcriptome data are ideally modeled via a negative binomial regression model that considers over-dispersion [Anders and Huber 2010], we plan to expand the choice of association tests currently incorporated in *CIF* methodologies. In addition, since these tests will rely on a regression framework, they can potentially be adjusted for confounding factors. In brief, the *CIF* framework provides generality and flexibility for enhancements in many contexts, including integrative multi-'omics' data analysis. In future work, our aim is to avoid a posteriori data integration (for instance fusing a methylome-transcriptome, genome-transcriptome, transcriptome-transcriptome networks via [Wang, et al. 2014]), but to develop a feasible *CIF*- based gene regulatory network inference method that can handle methylome, genome and transcriptome data as joint input to predict gene expression.

Conclusions

In this work, we investigated the performance and practical use of conditional inference trees (*CIT*s) and forest (*CIF*s) to infer gene regulatory networks from synthetic and real-life data. Synthetic data and data on model organisms were made available by the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project [Marbach, et al. 2010; Marbach, et al. 2009; Prill, et al. 2010]. Real-life data on type 1 diabetes was obtained from the GEO public repository (GEO #: GSE43488). We have shown that the conditional inference framework suggested by [Hothorn, et al. 2006] offers interesting possibilities for data integration, provided computational efficiency can be enhanced. In real-life settings of high-dimensional biological data, we recommend to use the CIF-based GRN inference approach based on conditional inference forests and node-specific *p*-values, adjusted for multiplicity in node selection by Monte Carlo resampling. In addition, we recommend randomly selecting about 1/3rd of the input variables at each node in the forest. Although more computationally intensive, this approach is less dependent on the number of randomly selected variables at each node than conditional inference trees with Bonferroni corrected *p*-values. Averaging node-specific *p*-values over trees in *CIF* ensembles and using these as variable importance scores to weight network edges, greatly facilitates construction of weighted networks such as GRNs. Indeed, for classic forests-derived variable importance scores it is not obvious to set a threshold above which two nodes need to be connected in the network, unless ROC or PR curves are constructed based on gold standard and the optimal threshold is derived from those. The statistical framework that underpins *CIF*s naturally leads to setting an overall "significance" level, such as 0.01. The latter is important when working with real-life biological data, for which the truth is largely unknown. Adopting this strategy on microarray gene-expression data for 121 type 1 diabetes patients and 1617 genes gave meaningful results, supported by the literature.

Acknowledgements

This research was funded by the Fonds de la Recherche Scientifique (FNRS), in particular F.N.R.S. research grant n° 22333518 (KB) and research project convention n° 2.4609.11 (KVS, KB). In addition, this paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

We thank Pierre Geurts and Benjamin Dizier from the Systems Modeling Unit, Faculty of Applied Sciences, Université de Liège (Belgium) for constructive comments and suggestions throughout this project. Special thanks to Patrick E. Meyer from the Bioinformatics and Systems Biology Unit, Faculty of Sciences, Université de Liège (Belgium) for providing valuable feedback to earlier versions of this manuscript.

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome biol 11(10):R106.

Baitaluk M, Kozhenkov S, Ponomarenko J. 2012. An integrative approach to inferring gene regulatory module networks. PLoS One 7(12):e52836.

Boulesteix AL, Janitza S, Hapfelmeier A, Van Steen K, Strobl C. 2015. Letter to the Editor: On the term 'interaction' and related phrases in the literature on Random Forests. Brief Bioinform 16(2):338-45.

Boulesteix AL, Janitza S, Kruppa J, König IR. 2012a. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2(6):493-507.

Boulesteix AL, Janitza S, Kruppa J, König IR. 2012b. Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics. University of Munich. Report nr 129.

Breiman L. 2001. Random forests. Machine learning 45(1):5-32.

Cho DY, Kim YA, Przytycka TM. 2012. Chapter 5: Network biology approach to complex diseases. PLoS Comput Biol 8(12):e1002820.

Davidson E, Levin M. 2005. Gene regulatory networks. Proceedings of the National Academy of Sciences of the United States of America 102(14):4935-4935.

Davis J, Goadrich M. The relationship between precision-recall and ROC curves; 2006; Pittsburgh, PA.

Essaghir A, Toffalini F, Knoops L, Kallin A, van Helden J, Demoulin JB. 2010. Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. Nucleic Acids Res 38(11):e120.

Hardiman G. 2004. Microarray platforms-comparisons and contrasts. Pharmacogenomics 5(5):487-502.

Horvath S, Dong J. 2008. Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol 4(8):e1000117.

Hothorn T, Hornik K, Strobl C, Zeileis A, Hothorn MT. 2014. Package ‘party’. Package Reference Manual for Party Version 0.9-998 16:37.

Hothorn T, Hornik K, Zeileis A. 2006. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics 15(3):651-674.

Hu Z, Killion PJ, Iyer VR. 2007. Genetic reconstruction of a functional transcriptional regulatory network. Nature genetics 39(5):683-687.

Hulme MA, Wasserfall CH, Atkinson MA, Brusko TM. 2012. Central role for interleukin-2 in type 1 diabetes. Diabetes 61(1):14-22.

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. 2010. Inferring regulatory networks from expression data using tree-based methods. PLoS One 5(9).

Ideker T, Krogan NJ. 2012. Differential network biology. Mol Syst Biol 8:565.

Jiang C, Xuan Z, Zhao F, Zhang MQ. 2007. TRED: a transcriptional regulatory element database, new entries and other development. Nucleic Acids Res 35(Database issue):D137-40.

Johnstone IM, Titterington DM. 2009. Statistical challenges of high-dimensional data. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 367(1906):4237-4253.

Kallionpaa H, Elo LL, Laajala E, Mykkanen J, Ricano-Ponce I, Vaarma M, Laajala TD, Hyoty H, Ilonen J, Veijola R and others. 2014. Innate immune activity is detected prior to seroconversion in children with HLA-conferred type 1 diabetes susceptibility. Diabetes 63(7):2402-14.

Kolchanov NA, Ignatieva EV, Ananko EA, Podkolodnaya OA, Stepanenko IL, Merkulova TI, Pozdnyakov MA, Podkolodny NL, Naumochkin AN, Romashchenko AG. 2002. Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucleic Acids Res 30(1):312-7.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559.

Levy ED, Landry CR, Michnick SW. 2009. How Perfect Can Protein Interactomes Be?

Liaw A, Wiener M. 2002. Classification and Regression by randomForest. R news 2(3):18-22.

Lin YY, Liu TL, Fuh CS. 2011. Multiple kernel learning for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell 33(6):1147-60.

Loh WY. 2011. Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(1):14-23.

Maarleveld TR, Khandelwal RA, Olivier BG, Teusink B, Bruggeman FJ. 2013. Basic concepts and principles of stoichiometric modeling of metabolic networks. Biotechnol J 8(9):997-1008.

Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Consortium D, Kellis M, Collins JJ and others. 2012. Wisdom of crowds for robust gene network inference. Nat Methods 9(8):796-804.

Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G. 2010. Revealing strengths and weaknesses of methods for gene network inference. Proc Natl Acad Sci U S A 107(14):6286-91.

Marbach D, Schaffter T, Mattiussi C, Floreano D. 2009. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. J Comput Biol 16(2):229-39.

McKnight A, Woodman A, Parkkonen M, Patterson C, Savage D, Forsblom C, Pettigrew K, Sadlier D, Groop P, Maxwell A. 2009. Investigation of DNA polymorphisms in SMAD genes for genetic predisposition to diabetic nephropathy in patients with type 1 diabetes mellitus. Diabetologia 52(5):844-849.

Michoel T, De Smet R, Joshi A, Van de Peer Y, Marchal K. 2009. Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. BMC systems biology 3(1):49.

Portales-Casamar E, Arenillas D, Lim J, Swanson MI, Jiang S, McCallum A, Kirov S, Wasserman WW. 2009. The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. Nucleic Acids Res 37(Database issue):D54-60.

Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G. 2010. Towards a rigorous assessment of systems biology models: the DREAM3 challenges. PLoS One 5(2):e9202.

Sahni N, Yi S, Taipale M, Fuxman Bass JI, Coulombe-Huntington J, Yang F, Peng J, Weile J, Karras GI, Wang Y and others. 2015. Widespread macromolecular interaction perturbations in human genetic disorders. Cell 161(3):647-60.

Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, Weiss V, Solano-Lira H, Martinez-Flores I, Medina-Rivera A and others. 2013. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res 41(Database issue):D203-13.

Strasser H, Weber C. 1999. On the asymptotic theory of permutation statistics.

Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. 2008. Conditional variable importance for random forests. BMC Bioinformatics 9:307.

Strobl C, Boulesteix AL, Zeileis A, Hothorn T. 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8:25.

Strobl C, Hothorn T, Zeileis A. 2009. Party on! A New, Conditional Variable Importance Measure for Random Forests Available in the party Package. University of Munich.

Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. 2014. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 11(3):333-7.

Wang YX, Huang H. 2014. Review on statistical methods for gene network reconstruction using expression data. J Theor Biol.

Westfall PH. 1993. Resampling-based multiple testing: Examples and methods for p-value adjustment: John Wiley & Sons.

Yao F, Coquery J, Le Cao KA. 2012. Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets. BMC Bioinformatics 13:24.

Zhu X, Gerstein M, Snyder M. 2007. Getting connected: analysis and principles of biological networks. Genes Dev 21(9):1010-24.

Figure captions

Figure 1 - Gene regulatory network framework based on *CIT / CIF*

a) Given gene expression data for a number of subjects or individuals, consider iteratively each gene expression as output (Y^i - response) and remaining gene expression as input (X^{-i} - predictors).
b) Construct a conditional inference tree (*CIT*) or conditional inference forest (*CIF*) per input/output.
c) Predict. Per node, aggregate over all available tree(s) to obtain a variable importance measure (VIM_{global}). Construct a non-symmetric adjacency matrix and hence a directed network.
d) Compare. The previously predicted network compare to a gold standard, whenever such a standard is available or use performance metrics such as area under the ROC curve (*AUROC*) or area under the precision-recall curve (*AUPR*). Edges that differ between gold standard and predicted network are shown in red.

Figure 2 - DREAM4 performance results – *mtry=k/3*
AUROC and *AUPR* expressed performance of considered GRN inference methodologies for each of the 5 DREAM4 networks included in the study and described in the methods section.

Figure 3 - DREAM4 performance results – variable *mtry*
Performance of proposed CIF_{mean} methods at various *mtry* values, assessed via the DREAM4 overall score. Overall scores are averaged over 5 networks.

Figure 4 – DREAM2 performance results – variable *mtry*
a) Performance of introduced CIF_{mean} methods based on the total area of *AUROC* and *AUPR*. b) A more detailed view of the *AUROC* and *AUPR* dynamics as a function of the *mtry* parameter.

Figure 5 – DREAM5 performance results – *mtry=k/3*
AUROC and *AUPR* expressed performance of considered GRN inference methodologies for each of the 3 DREAM5 networks included in the study and described in the methods section.

Figure 6 – DREAM5 performance results – variable *mtry*
Performance of CIF_{mean} methods at various *mtry* values, assessed based on DREAM5 overall score (average over 3 DREAM5 networks).

Figure 7 – The T1D Case study performance results – variable *mtry*
a) Performance of CIF_{mean} methods based on the total area of *AUROC* and *AUPR*. b) A more detailed view of the *AUROC* and *AUPR* dynamics as a function of the *mtry* parameter.

Supplementary material

Figure S1 – DREAM 4 performance results - *mtry=k/3*
GRN inference performance levels across 8 methodologies described in the methods section. Performance is quantified via the DREAM4 overall score as defined in for instance [Marbach, et al. 2012].

Figure S2 – DREAM2 performance results - *mtry=k/3*
Performance of CIF_{mean} and *RF* methods based on the total area of *AUROC* and *AUPR (mtry=k/3)*.

Figure S3 – DREAM5 performance results - *mtry=k/3*
GRN inference performance levels across CIF_{mean} methodologies. Performance is quantified via the DREAM5 overall score as defined in for instance [Marbach, et al. 2012].

Figure S4 – T1D Case study performance results - *mtry=k/3*
Performance of CIF_{mean} methods based on the total area of *AUROC* and *AUPR* (*mtry=k/3)*.

Figure S5 – DREAM4 GS networks
Network properties of the DREAM4 GS networks (1-5), each of size 100 (nodes).