Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original Paper

Genome analysis

Association rule mining to identify transcription factor interactions in genomic regions

Gaia Ceddia [1,*], Liuba Nausicaa Martino [2], Alice Parodi [2], Piercesare Secchi [2,3], Stefano Campaner [4] and Marco Masseroli [1]

[1] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, 20133, Italy
[2] MOX – Dipartimento di Matematica, Politecnico di Milano, Milan, 20133, Italy
[3] Center for Analysis, Decisions and Society, Human Technopole, Milano, Milan, 20157, Italy
[4] Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia (IIT), Milan, 20139, Italy

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Genome regulatory networks have different layers and ways to modulate cellular processes, such as cell differentiation, proliferation, and adaptation to external stimuli. Transcription factors and other chromatin-associated proteins act as combinatorial protein complexes that control gene transcription. Thus, identifying functional interaction networks among these proteins is a fundamental task to understand the genome regulation framework.

Results: We developed a novel approach to infer interactions among transcription factors in user-selected genomic regions, by combining the computation of association rules and of a novel *Importance Index* on ChIP-seq data sets. The hallmark of our method is the definition of the *Importance Index*, which provides a relevance measure of the interaction among transcription factors found associated in the computed rules. Examples on synthetic data explain the index use and potential. A straightforward pre-processing pipeline enables the easy extraction of input data for our approach from any set of ChIP-seq experiments. Applications on ENCODE ChIP-seq data prove that our approach can reliably detect interactions between transcription factors, including known interactions that validate our approach.

Availability: A R/Bioconductor package implementing our association rules and *Importance Index* based method is available at http://bioconductor.org/packages/release/bioc/html/TFARM.html

Contact: gaia.ceddia@polimi.it

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

1 Introduction

Transcription factors are regulatory proteins whose co-occurrence on genomic regions and interaction with each other can lead to the regulation of gene expression (Diamond *et al.*, 1990; Lundberg *et al.*, 2016). Gene regulation induced by transcription factors depends on their ability to bind DNA elements and to form protein complexes. Computational prediction of transcription factor interactions can hence provide insights on the genome regulatory framework.

Several algorithms have been proposed to infer protein-protein associations (Wixon, 2001), including transcription factor associations; some of them are based on co-expression data (Szklarczyk *et al.*, 2015), others on Bayesian methods that integrate different types of genomic data (McDowall *et al.*, 2009; Schmitt *et al.*, 2014), others rely on structure similarity and evolutionary conservation (Keskin *et al.*, 2008). Additionally, methods directly targeting interactions in chromatin have been proposed to better understand regulatory networks (Lundberg *et al.*, 2016). Bayesian networks (van Steensel *et al.*, 2010) and Markov random fields (Zhou and Troyanskaya, 2014) were first adopted to solve this task on a limited number of chromatin immunoprecipitation sequencing (ChIP-seq) data. Then, the ChromNet approach (Lundberg *et al.*, 2016) made it feasible on more than one thousand ChIP-seq experiments genome-wide. ChromNet is based on the computation of the inverse correlation matrix from ChIP-seq datasets of transcription factors and other chromatin-associated proteins. It then improves the inverse correlation network by expressing correlation relationships among groups of regulatory factors, as well as individual factors. Its relevance, in addition to the proposed method, is mainly in the provided results, made publicly available via Web interface (http://chromnet.cs.washington.edu/).

Association rule mining is a well-known technique in data mining and knowledge discovery. It is used to discover association rules in large transaction databases (Sun and Bai, 2008), which can unveil hidden relationships among frequent patterns of items present in a database (Datta *et al.*, 2016). Association rules have been used in many different applications, including marketing, text mining, and classification (Sun and Bai, 2008). Several examples of association rule mining exist in bioinformatics (Naulaerts *et al.*, 2013), such as to identify hotspots in cancer data (Agrawal and Choudhary, 2011), or to identify relevant genes in gene expression and methylation data (Mallik *et al.*, 2015). Many attempts have been undertaken to rank association rules or to weight them (Sun and Bai, 2008; Datta *et al.*, 2016; Mallik *et al.*, 2015); however, no importance measure of an item in a rule or rule set has been proposed yet.

Here, we innovatively use association rule mining to identify transcription factors' regulatory networks on chromatin. Our approach is based on the definition of a novel *Importance Index* to measure the relevance of each item in an association rule; this provides a ranking of each chromatin-associated factor's significance in the interaction with another specific factor, i.e., of most likely interactors of a particular factor. Compared to other methods, the advantages of our approach are the lower computational cost and the ability to evaluate transcription factor associations in user-selected genomic regions.

2 Methods

In this Section, we discuss the *Apriori* algorithm to find association rules, and we propose the novel *Importance Index* to identify interactors of a target item in the rules, where the target item is a transcription factor (TF) selected by the user to be studied and the rule items are other transcription factors. Moreover, we defined itemsets as the promoter regulatory regions with at least one TF binding site. In this setting, the *Importance Index* gives a measure of confidence that two or more transcription factors form a transcription factor complex with the selected TF within the promoter regulatory regions.

2.1 Apriori algorithm

Apriori is the algorithm most frequently used for the search of association rules (Agrawal and Srikant, 1994). Consider a set K = {k_1, k_2, ..., k_n} of n binary attributes, called *items*, and a set T = {t_1, t_2, ..., t_m} of m transactions, or *itemsets*, with every transaction t_i in T having a unique ID and consisting of a subset of items in K. An *association rule* over the dataset T is defined as an implication of the form:

$$X \rightarrow Y$$

where X and Y are two sets of items, respectively called *antecedent* (or left-hand-side, LHS) and *consequent* (or right-hand-side, RHS) of the rule, with $X, Y \subseteq K$ and $X \cap Y = \emptyset$.

An association rule is typically described by three measures: *support*, *confidence* and *lift*, representing the significance and interest of the rule.

- The *support* measures the rule frequency in the dataset, defined as:

$$\textit{supp}(X \rightarrow Y) = \frac{\textit{supp}(X \cup Y)}{|T|} \tag{1}$$

where $|T|$ is the number of transactions in T, $X \cup Y$ is a set of items in K and $supp(X \cup Y)$ is the support of the itemset $X \cup Y$ in T, with the support of X in T defined as:

$$\textit{supp}(X) = \frac{|\{t_i \in T : X \subseteq t_i\}|}{|T|} \tag{2}$$

that is the proportion of transactions t_i in the dataset T that contain the itemset X. The support of an association rule measures the frequency of the items satisfying the rule in the dataset and varies in the interval [0,1].

- The *confidence* is defined as:

$$\textit{conf}(X \rightarrow Y) = \frac{\textit{supp}(X \cup Y)}{\textit{supp (X)}} \tag{3}$$

It gives an estimate of the conditioned probability $P(Y|X)$, that is the probability of finding the RHS of the rule (i.e., the itemset Y) in the set T of transactions, given that such transactions also contain the LHS of the rule (i.e., the itemset X). Therefore, it measures the reliability of the inference made by the rule $X \rightarrow Y$. The higher is the confidence of the rule, the higher is the probability of finding the itemset Y in a transaction containing the itemset X. It varies in the interval [0,1].

- The *lift* is defined as:

$$\textit{lift}(X \rightarrow Y) = \frac{\textit{supp}(X \cup Y)}{\textit{supp (X)} * \textit{supp (Y)}} \tag{4}$$

It can be interpreted as the variation of the support of the rule with respect to the support obtained assuming that the rule LHS and RHS are independent. Therefore, it measures the strength of the rule, and varies in the interval $[0, \infty]$.

The general process of the Apriori algorithm consists of two steps: the frequent itemset generation and the rule generation (Agrawal and Srikant, 1994). The first step is to generate a frequency table of all items that occur in the transaction set, and to find the set of frequent items (i.e., of the items that have support greater than a threshold), which has k significant items (with $k \leq n$, where n is the total number of items in the K set). Then, starting from the most frequent item, the algorithm expands the search of frequent sets to all the possible pairs of significant items (i.e., both belonging to the set of frequent items) and applies the support threshold. This step is repeated for triplets of items, quartets, and so on. In other words, the frequent itemset generation starts from k-itemset (set of frequent items in the transaction set), adds one item at a time for each k-itemset (until $k = n - 1$), and computes the support. The new itemset has to pass the support threshold. The second step creates rules from each frequent itemset that satisfy the minimum confidence requirement. If a frequent itemset contains k elements, then the number of candidate association rules is equal to $2^k - 2$ (Agrawal and Srikant, 1994).

2.2 Importance Index

We aim to identify the most important items in the LHS itemsets generated by the Apriori algorithm taking into account the RHS as the target item(s) selected by the user. Since the number of association rules and LHS items in each rule can be very high, the entire list of LHS items in the whole set of rules alone does not provide an intelligible result without a measure of how much each item contributes to the existence of a rule. For example, let us consider the rule:

$$\{k_1 = 1, k_2 = 1, k_3 = 1\} \rightarrow \{k_t = 1\}$$

Just by looking at it, an analyst could not tell how much the presence of each item k_1, k_2 and k_3 contributes to the prediction of the k_t's presence. To find it out, we propose to substitute alternatively the presence of k_1, k_2 and k_3 in the rule with their absence, and in each case evaluate: if the modified itemset keeps existing in the dataset, and how the three quality measures of support, confidence and lift of the correspondent rule change. If the quality measures of a rule are equal to 0 when an item is set as absent in the pattern identified by the LHS of the rule (e.g., $\{k_1 = 0, k_2 = 1, k_3 = 1\}$ for the absence of the item k_1), then the presence of that item in the pattern $\{k_1 = 1, k_2 = 1, k_3 = 1\}$ is fundamental for the existence of the association rule $\{k_1 = 1, k_2 = 1, k_3 = 1\} \rightarrow \{k_t = 1\}$. Otherwise, if the modified rule keeps existing as relevant (i.e., support, confidence and lift are different from 0), and its quality measures are equal to the ones of the rule initially considered, then the presence of that item in the pattern $\{k_1 = 1, k_2 = 1, k_3 = 1\}$ is not fundamental for the existence of the association rule $\{k_1 = 1, k_2 = 1, k_3 = 1\} \rightarrow \{k_t = 1\}$.

The first step to evaluate the importance of items in the LHS is to choose the target item(s) k_t as the RHS and use the Apriori algorithm to compute significant association rules. Then, for each item $\hat{k}$ in the computed LHSs, we extract the subset of all relevant associations containing $\hat{k}$ in the LHS, named $R^{\hat{k}}$ (with $J = |\{R^{\hat{k}}\}|$ the number of rules in $R^{\hat{k}}$). Each element of $\{R_j^{\hat{k}}\}_{j=1:J}$ is described by a set of quality measures of support, confidence, and lift: $\{s_j^{\hat{k}}, c_j^{\hat{k}}, l_j^{\hat{k}}\}_{j=1:J}$.

Let $\{R_j^{\hat{k}-}\}_{j=1:J}$ be the set of rules obtained substituting the presence of the item $\hat{k}$ with its absence in each element of $\{R_j^{\hat{k}}\}_{j=1:J}$. For example, if $\hat{k}$ is k_1 and $R_j^{\hat{k}}$ is the rule $\{k_1 = 1, k_2 = 1, k_3 = 1\} \rightarrow \{k_t = 1\}$ with measures $\{s_j^{\hat{k}}, c_j^{\hat{k}}, l_j^{\hat{k}}\}$, then $R_j^{\hat{k}-}$ is the rule $\{k_1 = 0, k_2 = 1, k_3 = 1\} \rightarrow \{k_t = 1\}$ with measures $\{s_j^{\hat{k}-}, c_j^{\hat{k}-}, l_j^{\hat{k}-}\}$.

To analyze the importance of an item $\hat{k}$, we compare the two distributions $\{s_j^{\hat{k}}, c_j^{\hat{k}}\}_{j=1:J}$ and $\{s_j^{\hat{k}-}, c_j^{\hat{k}-}\}_{j=1:J}$ for each j in $\{1, ..., J\}$. We do not consider the lift since it is directly proportional to the confidence measure. We define the *Importance Index* of the item $\hat{k}$ in the rule $R_j^{\hat{k}}$ as:

$$imp(\hat{k})_j = \Delta s_j + \Delta c_j \tag{5}$$

with:

$$\Delta s_j = s_j^{\hat{k}} - s_j^{\hat{k}-}, \Delta c_j = c_j^{\hat{k}} - c_j^{\hat{k}-} \tag{6}$$

The importance of $\hat{k}$ in its set of rules $R^{\hat{k}}$ is obtained evaluating the mean of all its importances $imp(\hat{k})_j$ in the set of rules:

$$imp(\hat{k}) = \frac{\sum_{j=1}^{J} imp(\hat{k})_j}{J} = \frac{\sum_{j=1}^{J} \Delta s_j + \Delta c_j}{J} = \overline{\Delta s} + \overline{\Delta c} \tag{7}$$

where $\overline{\Delta s}$ and $\overline{\Delta c}$ are the means of the support and confidence variations over the total number of rules in which $\hat{k}$ is in the LHS and the target item(s) is set as the RHS. In Equation 6, variations are defined as the differences between two conditions: item $\hat{k}$'s presence and item $\hat{k}$'s absence for each rule. Δs varies in (-1,1] range, where $\Delta s = -1$ means that the item $\hat{k}$ and the target item(s) do not appear together (not possible by definition of $\hat{k}$'s association rules) and $\Delta s = 1$ means that the item $\hat{k}$ and the target item(s) appear together in every itemset. Also Δc varies in (-1,1] range; it is equal to -1 when item $\hat{k}$ does not occur with the target item(s) and it is equal to 1 when the frequency of $\hat{k}$'s association rules is the same as the frequency of k_t. Consequently, the *Importance Index* varies in (-2,2] range, reaching the lowest value when the frequency of the item $\hat{k}$'s absence is greater than the frequency of the item $\hat{k}$'s presence and the co-occurrence of item $\hat{k}$ and the target item(s) is equal to 0. It reaches the highest value when the item $\hat{k}$ and the target item(s) appear together in every itemset, or when the number of presences of the item $\hat{k}$ in the dataset are close to the total number of itemsets. Evaluating the index $imp(\hat{k})$ for each item $\hat{k}$ in the relevant association rules extracted allows ranking the items by their importance when the target item(s) is set as RHS. Items with high mean *Importance Index* are assumed fundamental for the existence of association rules; conversely, items with low mean importance do not significantly influence the pattern of items associated with the target item(s).

The *Importance Index* can be easily evaluated also on item couples, triplets, etc., by substituting the item $\hat{k}$ with a set of items (e.g., $\hat{k} = \{k_1, k_2\}$) and using the same procedure. Thus, we identify as $R^{\hat{k}}$ the set of rules containing both k_1 and k_2, and as $R^{\hat{k}-}$ the set of correspondent rules without the two items. This approach allows the identification of interactions between items that would be unrevealed just by looking at the list of association rules. When items represent transcription factors, this allows identifying and quantifying the contribution of one or more transcription factors to the co-factor complex of a given target factor.

2.3 Implementation

Our method is implemented as an open source software package, named TFARM, within the R/Bioconductor framework (Gentleman *et al.*, 2004). After the peer review process, TFARM is available in the Bioconductor release (http://bioconductor.org/packages/release/bioc/html/TFARM.html). From January 2018 to July 2019 it had 2,032 downloads by 964 distinct IPs. TFARM computes association rules and searches for high ranked transcription factors given a matrix of factor itemsets from ChIP-seq data, a target transcription factor, and minimum support and confidence thresholds as inputs. Visualization tools for the evaluation of the *Importance Index*, examples, and guidelines are provided in the package.

3 Results

In this section, we first provide two examples on synthetic simulated data to illustrate how our proposed *Importance Index* can discriminate the relevance of items in a dataset of itemsets. Then, we apply the *Importance Index* approach to real datasets of transcription factors to show the usefulness of our results and compare them with ChromNet's ones.

3.1 Example 1: Small synthetic dataset

Table 1 reports a small synthetic dataset of binary items organized in the form of a matrix, named *matrix of presences*, where each row describes an itemset, each column is a different item, and elements equal to 1 define the presences of an item in an itemset, whereas elements equal to 0 identify the absences of an item. Such a matrix can hence represent the presence/absence of bindings of transcription factors (items) in selected genomic regions (itemsets). To compute the importance of each item associated with an item of interest (or target item) we use our proposed *Importance Index*, which expresses how much the presences of an item are related to the presences of the target item. After setting the presence of the target item (e.g., item A) as the RHS of the association rules to search for, the Apriori method is applied to the matrix of presences in Table 1. Association rules with support threshold greater than zero are reported in Table 2. The influence of each item on the target item's presence is first computed as an *Importance Index* in each rule including the item, and then calculated as mean of these *Importance Indexes* over all item-associated rules. For example, the *Importance Index* of item B in the $\{B = 1\} \rightarrow \{A = 1\}$ rule is the sum of the support and

Table 1. Example 1 dataset.

Itemset ID	A	B	C	D	E
1	1	1	1	0	1
2	1	1	0	1	0
3	1	1	0	0	0
4	1	1	0	0	0
5	1	1	1	0	0
6	0	0	1	1	0
7	0	0	1	1	0

A, B, C, D and E are binary items in the dataset.

Table 2. Item association rules with A as target item and computed for the dataset in Table 1.

Rule ID	LHS	RHS	Support	Confidence
1	$E = 1$	$A = 1$	0.14	1.00
2	$D = 1$	$A = 1$	0.14	0.33
3	$C = 1$	$A = 1$	0.29	0.50
4	$B = 1$	$A = 1$	0.71	1.00
5	$C = 1, E = 1$	$A = 1$	0.14	1.00
6	$B = 1, E = 1$	$A = 1$	0.14	1.00
7	$B = 1, D = 1$	$A = 1$	0.14	1.00
8	$B = 1, C = 1$	$A = 1$	0.29	1.00
9	$B = 1, C = 1, E = 1$	$A = 1$	0.14	1.00

LHS is the left-hand-side and RHS is the right-hand-side of the rules.

confidence variations with respect to the same rule with B’s absence (i.e., $\{B = 0\} \to \{A = 1\}$), as in Equation 5. Overall, the *Importance Index* of B is the average of all *Importance Indexes* of B in the set of rules where B is present (B appears in five rules, as shown in Table 2). Table 3 shows the ranking of the items in the dataset of Table 1, according to their *Importance Index* and considering A as the target item. The ranking explains the behaviour of the items in the dataset with respect to the target item. Item B scores the highest *Importance Index* (Imp) and average variations of support ($\overline{\Delta s}$) and confidence ($\overline{\Delta c}$) over the five rules in which it is present (N° rules). Low values of the $\overline{\Delta s}$ for all items depend on the fact that no item in Example 1 dataset reaches a number of presences equal to the number of itemsets, i.e., the frequency of the item’s presences over the total number of itemsets is always smaller than 1. Conversely, $\overline{\Delta c}$ depends on the co-occurrences between the item and the target item. In the Example 1 dataset, B and A (the target item) always co-occur, C and A co-occur twice over the total four presences of C, E is present only once and together with A, and D has three presences, but only one co-occurs with the target item. Negative scores of $\overline{\Delta s}$ are reached when the absences of the item in the matrix are more numerous than its presences, as for D and E in the Example 1 dataset. Concerning the $\overline{\Delta c}$, the item B has the highest score in Table 3, since B and the target item are always and only present together. As defined in Equation 3, the confidence of a rule is proportional to the ratio between the frequency of the rule itemset and the frequency of the rule LHS. C and E have positive $\overline{\Delta c}$ because they are present in four rules, which appear more often than the rules with their absence. Instead, D is present in two rules; both of them continue to exist when computed for $D = 0$ and indeed their support and confidence measures are greater than the ones computed for $D = 1$. Thus, D has a negative $\overline{\Delta c}$. More on Example 1’s support and confidence distributions is in the Supplementary Material (Section S1 Figures S3-S6).

We defined the *Importance Index* of an item in an association rule as the unweighted linear combination of support and confidence variations, obtained substituting the presence of the item in the LHS of the rule with its absence (as in Equation 5). In this way, we assume that each of the two variations equally contributes to the evaluation of the item importance, and their sum gives enough information to rank the items in the LHS. Nevertheless, one of the two quality measures might be more (or less) sensitive than the other to the removal of the item from the rule, leading to a greater (or smaller) variation of one or both of the support and confidence values. We demonstrated that this is not the case, and our *Importance Index* definition is adequate, through the Principal Components Analysis (PCA) (Johnson and Wichern, 2007; Bro and Smilde, 2014) of the joint distributions of the couples ($\overline{\Delta s}$ and $\overline{\Delta c}$). The PCA identifies a sequence of linear subspaces capturing increasing proportions of the total variability of the data. Each subspace is defined in terms of an orthogonal basis whose elements are called Principal Components and are defined as linear combinations of the original variables, with coefficients called *loadings*. Figure 1 shows the proportion of total variance captured by each principal component found and its loadings. The first principal component explains 80% of the total variability of the delta measures, and it is identified by a linear combination of the two delta measures with equal weights, as the *Importance Index* defined in Equation 5. Other Principal Components Analyses performed on real ChIP-seq datasets are reported in the Supplementary Material Figures S12-S14.

Table 3. Item ranking for the dataset in Table 1, considering A as target item.

Ranking	Item	Imp	$\overline{\Delta s}$	$\overline{\Delta c}$	N° rules
1	B	1.28	0.28	1.00	5
2	C	0.38	0.00	0.38	4
3	E	0.04	-0.21	0.25	4
4	D	-0.76	-0.43	-0.33	2

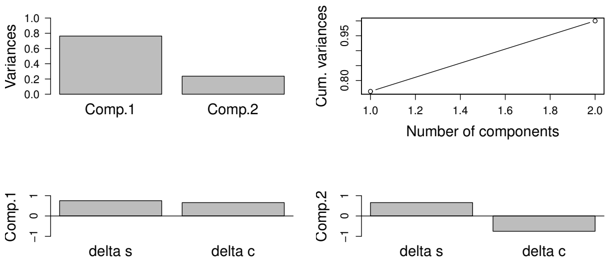

Fig. 1. Variances (upper left plot), cumulate variances (upper right plot) and loadings of the two principal components (lower plots).

3.2 Example 2: Larger synthetic dataset

Figure 2 shows the distribution of the presences of each item, and their relationship with the presence of the item A (set as target item), in a larger random synthetic, but realistic dataset with 7 items (e.g., transcription factors) and 1,000 itemsets (e.g., genomic regions of interest). Item A has 500 presences ($A = 1$). Items B, C and D have about 300 presences randomly distributed together with item A; they also appear in about 200

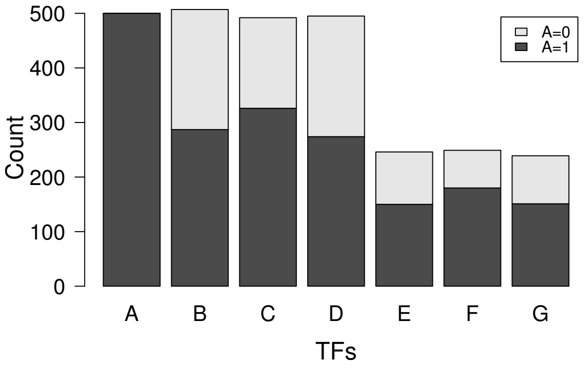

Fig. 2. Histogram of the presences of each item in the Example 2 dataset. $A = 0$: no co-occurrence with item A; $A = 1$: co-occurrence with item A.

itemsets where A is absent ($A = 0$). Items E, F and G have about 150 presences co-occurring and about 100 presences not co-occurring with item A. Once the target item is chosen, we can compute the *Importance Indexes* of all items in the dataset (items' association rules extracted are in Supplementary Material Table S1). Results for target item A are reported in Table 4. As expected, B, C, and D overcome E, F, and G in the item ranking thanks to their higher number of co-occurrences with A. Moreover, C has the maximum $\overline{\Delta s}$ and $\overline{\Delta c}$ since the frequency of rules where C and A co-occur exceeds the frequencies of such rules for B and D. For the same reason, F has greater $\overline{\Delta s}$ and $\overline{\Delta c}$ than G and E. Negative $\overline{\Delta s}$ scores are due to the excess of E, F, and G absences compared to their presences in the dataset. On the contrary, B, C and D have $\overline{\Delta s}$ values close to zero because the frequency of their presence is comparable to the frequency of their absence. More on Example 2's support and confidence distributions is in the Supplementary Material (Figures S9-S11). From these examples' evaluation we can define a high-ranked item as the one with high number of co-occurrences with the target item, such that the frequency of the rules with the item's presence in the LHS and the item's target presence in the RHS is close to the frequency of the item's presence in the dataset. Also, an optimal target item should be one of the most recurrent items in the dataset, such that its total number of presences is close to the number of itemsets (e.g., a transcription factor very frequent in the genomic regions of interest).

Table 4. Item ranking for the Example 2 dataset, considering A as target item.

Ranking	Item	Imp	$\overline{\Delta s}$	$\overline{\Delta c}$	N° rules
1	C	0.59	0.11	0.48	4
2	B	0.25	0.05	0.20	4
3	D	0.21	0.03	0.18	4
4	F	0.16	-0.14	0.30	1
5	G	-0.03	-0.20	0.17	1
6	E	-0.05	-0.20	0.15	1

Imp is the item *Importance Index*; $\overline{\Delta s}$ and $\overline{\Delta c}$ are the support and confidence mean contributions to the *Importance Index*.

3.3 Identification of transcription factor interactions

In this section, we show how the application of the proposed *Importance Index* to real transcription factor datasets allows the identification of relevant transcription factor interactions. First, we illustrate the datasets and explain the pipeline used to retrieve the matrix of presences for transcription factors in specific genomic regions; then, we analyze the importance of each evaluated transcription factor in the regulatory network of some selected transcription factors.

3.3.1 Datasets used

The analyses reported below have been performed on data retrieved from the Encyclopedia of DNA Elements (ENCODE) (The ENCODE Project Consortium, 2012). ChIP-seq narrow peak data regarding multiple cell lines were selected and extracted through GMQL (Masseroli *et al.*, 2018) to get binding locations of transcription factors in promoter regions. For the GMQL data pre-processing, two types of data have been considered:

1. hg19 human ENCODE data from ChIP-seq experiments including only *conservative* and *optimal* Irreproducible Discovery Rate (IDR) thresholded peaks
2. data concerning the localization of hg19 human promoter regions obtained extending the DNA coordinates of each Transcription Start Site (TSS) of known protein coding genes with an Entrez Gene ID, as available from the GENCODE repository version 10 (Frankish *et al.*, 2018), by 2,000 bases upstream and 1,000 bases downstream.

By mapping the enriched regions of each transcription factor on annotated promoters, we organized the data in binary matrices, one for each cell line considered. Each of these matrices represents a matrix of presences, built as follows: each row i represents a promoter region (or, in general, a genomic region of interest) and each column j, with $j > 4$, refers to the transcription factor j. The first four columns of the matrix specify the coordinates of the promoter regions: the chromosome ID, the first and last DNA base of the region, and the chromosome strand, respectively. So, the element (i, j) of the matrix of presences, with $j > 4$ is equal to 1 if the transcription factor j is present in the region i, or it is equal to 0 otherwise.

We considered the GM12878, HeLa-S3, and K562 cell lines, derived from human normal lymphoblastoid cells, cervical cancer immortalized cells, and chronic myelogenous leukemia immortalized cells, respectively. These lines were selected since, at the time of the analysis, they were the only ENCODE cell lines that, after the performed pre-processing, included ChIP-seq data for both MAX and MYC, as well as other transcription factors, which are discussed in the analysis described in Section 3.3.2. Their datasets respectively consist of 122, 55 and 210 transcription factors (i.e., columns of the matrices) and 16,758, 75,674 and 25,513 promoter regions (i.e., rows of the matrices) where at least one of the considered transcription factors is present (i.e., matrices do not include null rows) together with at least one of the H3K9ac or H3K4me3 histone modifications, which mark active promoters.

3.3.2 Identification of MAX interactors across cell lines

We focused our analysis on the evaluation of candidate interactors of MAX (MYC associated factor X), a protein-coding gene whose protein belongs to the basic helix-loop-helix leucine zipper (bHLHZ) family of transcription factors (Blackwood and Eisenman, 1991). MAX heterodimerizes with MYC as well as other MYC antagonists, such as MXI1, MAD, and MGA, to bind the DNA and regulate cell proliferation, differentiation, oncogenesis and other biological processes (Hurlin and Huang, 2006; Ewing *et al.*, 2007). These interactions are crucial for its activity. Thus, MAX behaves as a necessary cofactor for DNA binding and for most known biological activities of MYC and its antagonists (Hurlin and Huang, 2006). The analysis of MAX's interactors can, therefore, demonstrate the relevance and efficacy of our approach and give a real example of the *Importance Index* usage.

We applied our method to the matrices of presences built for the transcription factors of GM12878, HeLa-S3, and K562 cell lines, as described in Section 3.3.1. First, we chose MAX as the target transcription factor, and then we set the minimum confidence and support thresholds to find the association rules. For all cell lines, the minimum confidence threshold was set to 0.9, since it represents the probability of finding the LHS of the rule given the pattern in the RHS. Setting a lower confidence threshold would mean finding also less reliable rules, whereas setting a higher threshold would result in a limited number of rules for the analysis. The minimum support threshold was set in order to obtain a defined number of rules, 200 rules in this specific instance. In Section S2.1 of the Supplementary Material, we present guidelines for choosing the minimum support threshold and their statistical validation, which strongly supports the validity of our choice for the data sets used in this paper, and which we believe can be used as guidelines for similar data sets. For each cell line, we first computed the association rules with MAX's presence as the RHS, and then the *Importance Indexes* associated with each candidate interactor of MAX, building a ranking of such candidate interactors for each cell line (Supplementary Material Tables S5-S7). Figure 3 shows the results organized as a network, where each node is a transcription factor, each

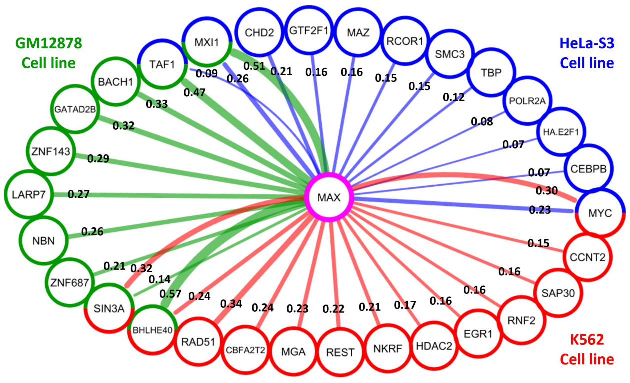

Fig. 3. MAX (pink node) interaction network for the GM12878 (green edges), HeLa-S3 (blue edges), and K562 (red edges) cell lines. Nodes represent transcription factors, and edges represent candidate interactions between MAX and other transcription factors
, based on their weight that is the *Importance Index*.

edge represents the *Importance Index*, and each cell line is encoded with a different color. MYC resulted one of the strongest MAX's interactors, confirming the importance of this well-known interaction in the literature (Ewing *et al.*, 2007). Conversely, MAX is the highest-score interactor of MYC in cell lines HeLa-S3 and K562 (Supplementary Material Tables S8-S9), confirming that MYC heterodimerizes with the transcription factor MAX to regulate the transcription of a large fraction of the genome (Ewing *et al.*, 2007). Even if the GM12878 dataset contains MYC data, MYC is not in the obtained GM12878 ranking due to its low presence in the considered GM12878 matrix (only 0.8% of presences on the evaluated promoter regions). Although we selected high-quality ChIP-seq experiments, the small number of MYC's occurrences in GM12878 (i.e., in only 87 out of 11,367 promoter regions) is probably due to the low sensibility of the ChIP during this specific GM12878 experiment, which makes it unreliable. Other high-ranked interactors of MAX resulted MXI1 and MGA; both are well-known MAX's interactors (Hurlin and Huang, 2006).

To comprehensively evaluate the obtained results, we compared them with both the known MAX's interactors in the BioGRID protein-protein interaction database (Stark *et al.*, 2006), and with the computational results of ChromNet (Lundberg *et al.*, 2016). The latter one is a popular statistical method and tool proposed to infer interaction networks of regulatory factors genome-wide, based on the conditional dependency between transcription factors and groups of transcription factors. Supplementary Material Section S2.2.1, Figures S18-S20, and Tables S10-S12 report the comparison results for each considered cell line. Overall, out of the 28 transcription factors that we found interacting with MAX, 20 were also in the input dataset evaluated by ChromNet. Of these latter ones, 18 (90%) were also found by ChromNet, and 6 (30%) of them are also known as MAX's interactors in BioGRID. The remaining 2 (10%) transcription factors found as candidate interactors of MAX were BACH1 and ZNF143 (the latter one with a lower Importance Index). BACH1 in BioGRID is known to interact with NRF2, which is in the interactors' network of MYC; this could support the possible interaction between BACH1 and MAX. Among the 8 transcription factors we found as candidate interactors of MAX that were not in the input dataset of ChromNet, we identified MGA, a well-known MAX's interactor in BioGRID. Comparison results confirm the relevance of our novel approach and accuracy of its results, whose assessment outperforms the corresponding ChromNet results. In fact, out of the 30 transcription factors that ChromNet found interacting with MAX, 18 (60%) were also found by our method and 6 (20%) of them are also known MAX's interactors in BioGRID. None of the remaining 12 (40%) transcription factors that only ChromNet found as candidate interactors of MAX are confirmed in BioGRID.

We also evaluated candidate interactors of MAX in random genomic regions (i.e., control regions). To this aim, we performed the *Importance Index* approach on randomized regions of interest, and we compared the results with candidate interactors of MAX in the promoter regions (Supplementary Material Section S2.2.2). As expected, comparisons show lower presences of factors in the control regions and different candidate interactors (Figures S21-S25).

3.3.3 Identification of multiple factor interactions: an example case

MXI1, RAD21, and SMC3 co-localize in HeLa-S3 cervical carcinoma cells (Lundberg *et al.*, 2016); this would suggest that they interact with each other. However, only the RAD21-SMC3 and SMC3-MXI1 interactions are described in the BioGRID database (Gupta *et al.*, 1998; Huttlin *et al.*, 2017); conversely, the RAD21-MXI1 interaction is not present in BioGRID. Moreover, an extensive study on RAD21 interactions (Panigrahi *et al.*, 2012) revealed more than 200 interactors, but among them, it did not identify MXI1. In our HeLa-S3 matrix of presences, RAD21, MXI1, and SMC3 are present in 1584, 4846 and 3756 promoter regions, respectively. MXI1 and RAD21 co-localize with each other in 279 regions, where also SMC3 is present, leading to the hypothesis that the RAD21-MXI1 pair can interact with SMC3. *Importance Index* based rankings of RAD21, MXI1, and SMC3 candidate interactors reveal that RAD21 and MXI1 are not associated with each other due to their low number co-occurrences (Supplementary Material Section S2.3 and Tables S13-S16).

To prove that our method can detect interactions among more than two regulatory factors, in particular among RAD21, MXI1 and SMC3, we evaluated the *Importance Indexes* of transcription factor pairs given SMC3 as target transcription factor. The obtained *Importance Indexes*' matrix is shown as a heatmap in Figure 4, where the lowest *Importance Indexes* are in white and the highest ones are in black. Figure 4 shows that the RAD21-MXI1 pair has one of the highest scores (see also Supplementary Material Table S16), indicating that the co-presence of RAD21 and MXI1 implies the presence of SMC3, i.e., RAD21, MXI1, and SMC3 are all associated together. The heatmap shows also that RAD21 and CTCF, paired with any of the other factors, are most likely to interact with SMC3. Conversely, MXI1 has generally low *Importance Indexes* because it co-localizes poorly with other regulatory factors in the heatmap with respect to its total number of presences in the promoter regions. Yet, it reaches one of the highest values of the *Importance Index* matrix in association with RAD21, i.e.,

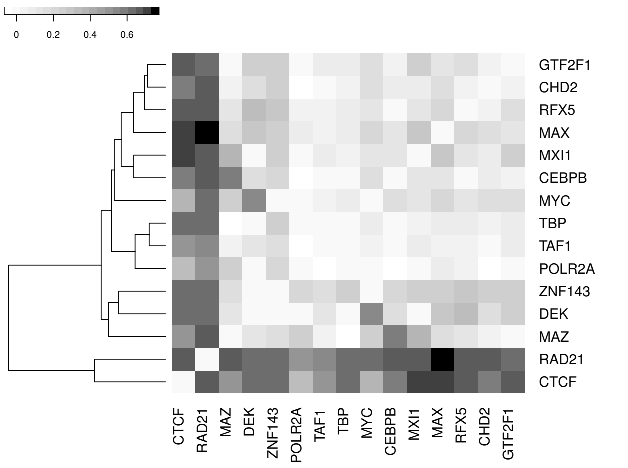

Fig. 4. Mean Importance Indexes of SMC3 candidate interactor pairs in HeLa-S3 cell line.

RAD21-MXI1 pair co-localizes more often with SMC3 than other pairs in HeLa-S3 cell line.

4 Discussion and Conclusions

Interactions among transcription factors strongly influence gene regulation, and several attempts have been made to model and understand regulatory factors’ networks. To compute how much a transcription factor contributes to the existence of a certain complex of transcription factors we propose a novel *Importance Index*, based on the combination of quality measures of association rules used to find possible transcription factor associations, where rules’ general items are transcription factors and the target item is one or more transcription factor(s) selected by the user.

The use of association rules and definition of the novel *Importance Index* allowed the development of our efficient interaction-detecting algorithm for the construction of regulatory networks. It is designed for the analysis on user-selected genomic regions and easy to reproduce; the matrices of presences needed as input can be easily built from ChIP-seq experiment data, and its implementation is publicly available as an R/Bioconductor package. Furthermore, it is computationally inexpensive; experiments run on an MS-Windows machine equipped with an Intel i7-8750H processor and 16 GB of RAM required only 5, 14, and 8 minutes to compute the *Importance Indexes* for the GM12878, HeLa-S3, and K562 cell lines, respectively, with the use of 3.0 GB of RAM for each computation.

Comparison of our results with known interactions in the BioGRID database and with ChromNet’s inferred networks demonstrate the ability of our approach to detect reliable interactions. We note that our and ChromNet’s methods have different objectives: ChromNet considers the whole genome globally and processes ChIP-seq reads aligned to the reference genome, involving numerous data and a computationally expensive processing. Conversely, our approach considers ChIP-seq binding enriched regions and allows quick and ad hoc analyses with results comparable to and even better than ChromNet’s ones, and confirmed by the literature. Thus, the *Importance Index* method gives local information about interactors of a target transcription factor, i.e., it can be reliably used when the user searches for interactions in specific genomic regions of interest; whereas, ChromNet provides genome-wide-based evaluations of a transcription factor network.

Future work will be focused on developing an ensemble approach including the *Importance Index* and other algorithms to infer a full transcriptional regulatory network. Thus, the aim will be assembling transcription factor and gene associations to evaluate if a predicted TF complex is regulating a target gene.

Acknowledgements

We thank Stefano Perna from the GeCo group of Politecnico di Milano for useful discussions and insights.

Funding

This work was supported by the ERC Advanced Grant 693174 "Data-Driven Genomic Computing (GeCo)" and by a grant of the Italian Association for Cancer Research-AIRC to SC (IG 21663).

Conflict of Interest: none declared.

References

Agrawal, A. and Choudhary, A. (2011). Identifying hotspots in lung cancer data using association rule mining. In *Proc. ICDMW11*, pages 995–1002.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proc. VLDB94*, pages 487–499.

Blackwood, E. and Eisenman, R. (1991). Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. *Science*, **251**(4998), 1211–1217.

Bro, R. and Smilde, A. K. (2014). Principal component analysis. *Anal Methods*, **6**, 2812–2831.

Datta, S. *et al.* (2016). Mining and ranking association rules in support, confidence, correlation, and dissociation framework. In *Proc. FICTA16*, pages 141–152.

Diamond, M. *et al.* (1990). Transcription factor interactions: selectors of positive or negative regulation from a single DNA element. *Science*, **249**(4974), 1266–1272.

Ewing, R. M. *et al.* (2007). Large-scale mapping of human protein–protein interactions by mass spectrometry. *Mol Syst Biol*, **3**(1).

Frankish, A. *et al.* (2018). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*, **47**, D766–D773.

Gentleman, R. *et al.* (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol*, **5**, 80.

Gupta, K. *et al.* (1998). Mmip1: a novel leucine zipper protein that reverses the suppressive effects of Mad family members on c-myc. *Oncogene*, **16**, 1149–1159.

Hurlin, P. J. and Huang, J. (2006). The MAX-interacting transcription factor network. *Semin Cancer Biol*, **16**(4), 265–274.

Huttlin, E. *et al.* (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature*, **545**, 505–509.

Johnson, R. and Wichern, D. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall.

Keskin, O. *et al.* (2008). PRISM: protein-protein interaction prediction by structural matching. In *Methods Mol Biol*, volume 484, pages 505–521. Humana Press.

Lundberg, S. M. *et al.* (2016). ChromNet: Learning the human chromatin network from all ENCODE ChIP-seq data. *Genome Biol*, **17**(1), 82.

Mallik, S. *et al.* (2015). RANWAR: Rank-based Weighted Association Rule mining from gene expression and methylation data. *IEEE Trans Nanobioscience*, **14**(1), 59–66.

Masseroli, M. *et al.* (2018). Processing of big heterogeneous genomic datasets for tertiary analysis of Next Generation Sequencing data. *Bioinformatics*, page bty688.

McDowall, M. D. *et al.* (2009). PIPs: human protein–protein interaction prediction database. *Nucleic Acids Res*, **37**, D651–D656.

Naulaerts, S. *et al.* (2013). A primer to frequent itemset mining for bioinformatics. *Brief Bioinform*, **16**, 216–231.

Panigrahi, A. *et al.* (2012). A cohesin–RAD21 interactome. *Biochem J*, **442**(3), 661–670.

Schmitt, T. *et al.* (2014). FunCoup 3.0: database of genome-wide functional coupling networks. *Nucleic Acids Res*, **42**(D1), D380–D388.

Stark, C. *et al.* (2006). BioGRID: a general repository for interaction dataset. *Nucleic Acids Res*, **34**, D535–D539.

Sun, K. and Bai, F. (2008). Mining weighted association rules without preassigned weights. *IEEE Trans Knowl Data Eng*, **20**, 489–495.

Szklarczyk, D. *et al.* (2015). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*, **43**(D1), D447–D452.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(414), 57–74.

van Steensel, B. *et al.* (2010). Bayesian network analysis of targeting interactions in chromatin. *Genome Res*, **20**, 190–200.

Wixon, J. (2001). Website review: protein-protein interactions on the web. *Comp Funct Genomics*, **2**, 338–343.

Zhou, J. and Troyanskaya, O. (2014). Global quantitative modeling of chromatin factor interactions. *PLOS Comput Biol*, **10**(3), 1–13.