scientific reports

OPEN

Multi-task learning for the simultaneous reconstruction of the human and mouse gene regulatory networks

Paolo Mignone[1], Gianvito Pio[1]✉, Sašo Džeroski[2] & Michelangelo Ceci[1,2]

The reconstruction of Gene Regulatory Networks (GRNs) from gene expression data, supported by machine learning approaches, has received increasing attention in recent years. The task at hand is to identify regulatory links between genes in a network. However, existing methods often suffer when the number of labeled examples is low or when no negative examples are available. In this paper we propose a multi-task method that is able to simultaneously reconstruct the human and the mouse GRNs using the similarities between the two. This is done by exploiting, in a transfer learning approach, possible dependencies that may exist among them. Simultaneously, we solve the issues arising from the limited availability of examples of links by relying on a novel clustering-based approach, able to estimate the degree of certainty of unlabeled examples of links, so that they can be exploited during the training together with the labeled examples. Our experiments show that the proposed method can reconstruct both the human and the mouse GRNs more effectively compared to reconstructing each network separately. Moreover, it significantly outperforms three state-of-the-art transfer learning approaches that, analogously to our method, can exploit the knowledge coming from both organisms. Finally, a specific robustness analysis reveals that, even when the number of labeled examples is very low with respect to the number of unlabeled examples, the proposed method is almost always able to outperform its single-task counterpart.

Gene regulation is the process that allows a cell to express a particular group of genes and to inhibit others in specific contexts. For example, a nerve cell has the same genome as a muscle cell, but they are different because of the different sets of expressed genes in each of them. This explains how the cells of different tissues have different proteomes, that is, different sets of proteins produced as a result of the selective expression of a gene or a group of genes. Since tumor cells are mainly caused by the expression of genes outside the original context of the cell, the understanding of gene regulation mechanisms appears to be fundamental to study various forms of cancer[1,2]. In this context, the analysis of Gene Regulatory Networks (GRNs) appears to be a fundamental task.

A GRN represents the system of regulatory genes and their interactions that determine the genetic functions to be expressed in cells of each spatial domain in the organism, at every stage of development. This includes the expression of regulatory genes (i.e., genes encoding transcription factors), genes that encode intercellular signaling functions, and genes that participate in downstream differentiation and morphogenesis functions[3]. As stated by Smith et al.[4], identifying the structure of GRNs helps in the biological understanding of disease mechanisms and increases possibilities for better medical/clinical care by improving diagnostics, prognostics and treatment. In particular, the reconstruction of a GRN comprises the identification of pairwise interactions between genes (i.e., nodes in the network) that participate in the same biological processes or that perform together specific biological functions that shape a system's behavior and function[5].

There are several techniques to elucidate the structure of a gene regulatory network. Some examples include ChIP-chip or ChIP-sequencing[6], bacterial one-hybrid systems[7] or protein-binding microarrays[8]. However, the validation process is often technically demanding, expensive and time-consuming[9].

Alternatively, the task can be supported by computational approaches that analyze the expression levels of genes, measured under different conditions. Since the high availability of such data makes computational approaches affordable, there has been a significant increase in computational methods proposed in the literature[10–13]. The task at hand is also referred to as "reverse-engineering" or "gene network reconstruction".

[1]Department of Computer Science, University of Bari Aldo Moro, Bari 70125, Italy. [2]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana 1000, Slovenia. ✉email: gianvito.pio@uniba.it

Existing methods usually do not rely on a single theory, but on multiple classes of statistical/mathematical methods and information/machine learning theory. In this context, the Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenges have also contributed to the development of this task. In particular, in follow-up studies, it has been shown that combining multiple approaches[14,15] or multiple sources[16,17] can be beneficial for GRN reconstruction.

Considering Gene Network Reconstruction as a machine learning task, it can be formulated as a link prediction problem via binary classification, where existing relationships among genes (i.e., gene regulation activities that have already been validated in the laboratory) can be considered as the set of *positive examples*. On the other hand, pairs of genes, for which there is a confirmation about the non-existence of the regulation, can be considered as *negative examples*. However, validation efforts and resources are usually spent to prove the existence of gene interactions, rather than their non-existence. This means that all the possible gene pairs for which there is no web-lab validation cannot be considered negative examples, but rather *unlabeled examples*. This context makes the adoption of classical supervised machine learning methods inappropriate or even inapplicable, and requires the design of semi-supervised learning methods, which are also explicitly able to work in the absence of negative examples, i.e., in the *positive-unlabeled* setting[17]. This is the most challenging setting, especially considering that the number of available positive examples is usually significantly lower than the number of unlabeled examples.

In order to face the above challenges, in this paper we propose a machine learning method for gene network reconstruction, which works in the positive-unlabeled setting and alleviates the issues arising from the limited availability of labeled data. In particular, the method proposed in this paper relies on a *transfer learning* approach that is able to exploit the knowledge of a source domain D_s to improve the result of a task performed on the target domain D_t. In our case, the data in each domain represents the expression levels measured for genes of a different organism.

Methodologically, we propose a specific kind of transfer learning approach, namely, a *multi-task* method[18], whose main advantage is the ability to simultaneously solve the task on both domains, and possibly exploit dependencies between them that could lead to improved accuracy of reconstruction. In particular, we aim at simultaneously reconstructing the gene regulatory networks of two related organisms, namely, the human and the mouse regulatory networks, by considering a novel instance mapping which is guided by the notion of genetic orthology[19].

State-of-the-art supervised machine learning methods employ a training set of examples which represents a sample of the population under analysis, described by a feature vector and associated with a known target value. These methods learn a prediction model which is able to assign a target value to unseen examples. This approach is widely proven to be effective if the set of examples is large enough and if the dataset is completely labeled, i.e., each example has a label (target value).

Therefore, classical supervised machine learning algorithms can be naturally applied to solve the task of network reconstruction, where: (i) each example corresponds to a (possible) relationship between two genes; (ii) features correspond to expression data regarding the two genes; (iii) labels can be $\{Yes, No\}$, depending on whether the interaction exists or not, or a numerical value representing the degree of certainty of the interaction. However, the quality of the reconstruction can be affected by the poor availability of labeled examples. Moreover, in this specific application domain, the available examples are usually only positive, i.e., they are only examples of existing interactions (see, for example, the well-known database BioGRID[20], that contains only existing gene regulations, without any examples of verified non-existing ones).

In the literature, we can find different approaches to face this challenge, that usually work in the positive-unlabeled learning setting. They can be classified according to three categories[21]:

(a) *two-step methods*, that identify a set of negative examples from the set of unlabeled examples and then, in the second step, exploit off-the-shelf supervised learning methods to build the final predictive model[22–24];
(b) *instance-weighting methods*, that estimate the reliability of each unlabeled example and exploit it as a weight or a cost while learning the prediction model[25];
(c) *noisy negative methods*, that consider the unlabeled set of examples as highly noisy negative examples[22,26].

The method proposed in this paper partially falls in category (b), that, according to previous studies[15,27], allows us to avoid the imposition of strong assumptions on the negative examples, made by the methods in categories (a) and (c). However, as we explain in detail in "Methods" section, the estimated weight is used as a target value, rather than as a weight. Moreover, as introduced in "Introduction" section, the proposed method is based on a multi-task approach which simultaneously solves the network reconstruction task for two organisms, namely, human and mouse, possibly exploiting dependencies and similarities among them.

Since the method proposed in this paper solves the gene network reconstruction task by exploiting a *transfer learning* approach, specifically based on *multi-task learning*, in the following subsections, we provide some background notions and briefly review existing methods in these fields.

Transfer learning.

One possible solution to overcome the scarcity of labeled examples is the adoption of transfer learning approaches[18,28], that aim at exploiting the knowledge about another related domain D_s, called *source domain*, to improve the quality of the results on the main domain D_t, called *target domain*.

Formally, in a classical supervised learning setting, given (i) the feature space X of training examples, (ii) the output space Y, and (iii) n training examples $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, such that $x_i \in X$ and $y_i \in Y$, the goal is to learn a function $f: X \rightarrow Y$, that predicts the label/value of unseen, unlabeled examples.

Transfer learning differs from this formulation since it works on two different domains. Formally, given:

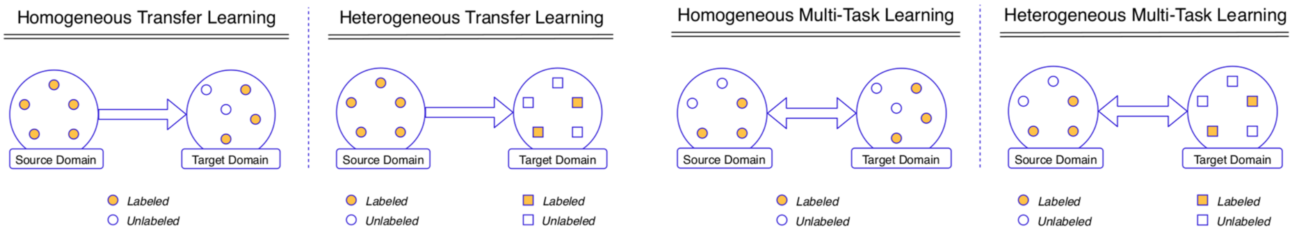

Figure 1. Homogeneous vs Heterogeneous transfer learning settings (left); Homogeneous vs Heterogeneous Multi-Task learning (right). The shape of the instances represents their feature space, while the arrow represents the direction of the transfer of knowledge between the domains.

- the source and the target feature spaces X_s and X_t;
- the output spaces Y_s and Y_t;
- n_s training examples $\{(x_1^s, y_1^s), (x_2^s, y_2^s), \ldots, (x_{n_s}^s, y_{n_s}^s)\}$ s.t. $x_i^s \in X_s$ and $y_i^s \in Y_s$ for the source domain;
- n_t training examples $\{(x_1^t, y_1^t), (x_2^t, y_2^t), \ldots, (x_{n_t}^t, y_{n_t}^t)\}$ s.t. $x_i^t \in X_t$ and $y_i^t \in Y_t$ for the target domain;

the goal is to learn a function f_t: $X_t \rightarrow Y_t$ on the target domain, also exploiting the knowledge acquired by learning a function f_s: $X_s \rightarrow Y_s$ on the source domain.

In the literature, we can find several transfer learning approaches, which were designed either as general-purpose frameworks or as specific methods, tailored for solving specific tasks of an application domain. Such approaches can be classified according to two categories (see Fig. 1 (left)):

- *homogeneous*, when the source and the target domains are described according to the same feature space (i.e., $X_s = X_t$);
- *heterogeneous*, where there are no restrictions on the feature spaces (i.e., $X_s \neq X_t$).

The heterogeneous setting is clearly more difficult to handle, since it is necessary to design a strategy to transform both the feature spaces into a common feature space, or to make them comparable. For example, some heterogeneous transfer learning approaches[29–31] assume that the source and the target domains are described with the same number of features and identify a shared feature subspace, where the difference between data distributions is minimized.

Another categorization of transfer learning methods[18] distinguishes among:

- *instance-based* methods, that usually perform a reweighing of the source domain instances, that are then directly used during the training for the target domain (see[16,32,33]);
- *parameter-based* methods, that aim to transfer the knowledge through some parameters shared by the models learned for the source and the target domains (see[17,34,35]);
- *feature-based* methods, that perform knowledge transfer by identifying a shared feature space (see[29–31,36,37]).

Focusing on transfer learning approaches proposed in the field of bioinformatics, in the literature we can find a recent work that aims to classify breast cancer tumors, as Estrogen-Receptor-positive (ER-positive) or Estrogen-Receptor-negative (ER-negative), by exploiting two different data sources[38]. A different approach, based on deep learning, has been used for molecular cancer classification[39], where the feature representation learned while classifying two tumor types also exploits information conveyed, in an unsupervised manner, by other tumor types. Breckels et al.[40] propose to extend a state-of-the-art transfer learning framework to solve the predictive task of mouse protein sub-cellular localization.

To the best of our knowledge, all the cited methods require a fully labeled training set, or assume the presence of some negative examples, following the strategies (a) or (c) described in "Introduction" section. In our previous work[16,17], we overcame this limitation by designing methods based on strategy (b), i.e., based on instance-weighting. In particular, these methods exploit the knowledge coming from the reconstruction of the mouse GRN for the reconstruction of the human GRN, in a homogeneous transfer learning setting. However, the main limitations of these methods are: (i) their inability to solve the gene network reconstruction task for both organisms simultaneously, and (ii) the homogeneous setting in which they work, that makes them hardly applicable if the gene relationships of the considered organisms are represented in different feature spaces.

The approach we propose in this paper exhibits the advantages of our previous work[16,17], without their limitations. In particular, the proposed method works in a multi-task learning setting, which aims at solving both gene network reconstruction tasks simultaneously, and which can analyze the considered organisms in either homogeneous or heterogeneous feature spaces. Moreover, according to the second categorization[18], our method falls in the category of *feature-based* transfer learning methods, since, as we will describe in "Methods" section, we identify a common feature space by exploiting the concept of genetic orthology.

Multi-task learning. A specific sub-category of transfer learning methods is represented by *multi-task learning* methods, which aim at simultaneously solving the task for both the source domain D_s and the target domain D_t. Such an advantage is not commonly present in standard transfer learning methods, which usually aim to facilitate or improve the task for the target domain only. On the contrary, multi-task learning approaches are able to optimize both tasks simultaneously, through multiple objective (or loss) functions, or their combination. The simultaneous consideration of the two tasks allows us to take into account possible bidirectional dependencies, which cannot be considered in single-task scenarios, even if a unidirectional transfer learning approach is applied multiple times.

Several complex machine learning applications have taken advantage of multi-task approaches, ranging from natural language processing[41] and speech recognition[42] to computer vision[43] and GRN reconstruction[44]. To the best of our knowledge, there is only one multi-task learning method in the literature that is able to work in a positive-unlabeled setting[45]. However, it requires that some of the solved tasks are classical supervised tasks, where the training set also includes negative examples. This makes its application inappropriate in our case, since, in principle, both the gene network reconstruction tasks are posed in the positive-unlabelled setting. This is an important aspect, as well as a strong contribution provided by our method. Indeed, although we can find several multi-task methods that are able to work in the semi-supervised setting (e.g.,[46,47]), they cannot be easily adapted to work in the positive-unlabeled setting for both the considered gene network reconstruction tasks, due to the inherent additional challenges introduced by the absence of negative examples. Therefore, our method simultaneously exhibits all the following characteristics:

- it can work with no negative examples, using positive and unlabelled examples of both domains (*positive-unlabelled*);
- it is able to transfer the knowledge acquired in the reconstruction of a GRN of an organism for the reconstruction of the GRN of another organism (*transfer learning*);
- it can simultaneously reconstruct (see Fig. 1 (right)) the GRN of two organisms, i.e., the knowledge is transferred bidirectionally (*multi-task learning*).

It is noteworthy that multi-task approaches are closely related to multi-target prediction methods. Indeed, multi-target prediction refers to the (possibly simultaneous) prediction of multiple variables of interest for the same units of observation[48]. In our case, we are interested in predicting the existence of relationships between genes of two different organisms. Therefore, considering an output variable for each organism leads the considered task to be conceptually close to a multi-target prediction task (since it is in fact a multi-target link prediction task). This aspect will be clearer in the next section, where we describe how we employ our multi-target prediction approach to solve the network reconstruction task for two organisms simultaneously.

Methods

In this section, we describe our method for simultaneous reconstruction of two GRNs in a multi-task learning setting. In particular, we will focus on the reconstruction of the human and mouse GRNs. To this end, we exploit the Predictive Clustering framework, that has proved its effectiveness in the presence of different forms of autocorrelation[49], i.e., when objects that are close to each other appear more related than distant ones. This is useful in the context of network data, like the GRNs under study, where genes close in the network are expected to show a similar behavior or to participate in the same biological processes.

In particular, we exploit the Predictive Clustering Tree (PCT) method implemented in the system CLUS[50]. CLUS is a decision tree and rule induction system that unifies unsupervised clustering and predictive modeling. It has been employed in several recent works to solve multi-target prediction tasks for a single domain. The approach proposed in this paper can be considered the first attempt to employ the PCT multi-target prediction method implemented in CLUS to work in a multi-task learning setting, where the variables to be predicted are associated with two different tasks.

Multi-target prediction methods are generally categorized according to whether they build multiple *local* models, i.e., one model for each target variable, separately, or one *global* model, i.e., a single predictive model that is able to predict the whole set of target variables simultaneously. Global models are generally more effective than their local counterparts[51], due to their ability to capture dependencies in both the input and the output spaces. In this paper we exploit the capability of CLUS to learn a global model, and we consider the degrees of existence of each gene interaction in the two organisms as two target variables of a multi-target regression task. This is achieved by representing the same examples of gene interactions, in the two organisms, in a common feature space. In particular, in order to find a match between the genes in the two organisms, we exploit the concept of orthologous genes, that are genes in different species that originated by vertical descent from a single gene of the last common ancestor.

Before describing in detail the proposed multi-task approach, we introduce some useful notions and formally define the problem that we solve. Let:

- G_h (resp., G_m) be the set of considered genes for the human (resp., mouse) organism;
- $B_h \subseteq G_h \times G_h$ (resp., $B_m \subseteq G_m \times G_m$) be the set of (biologically validated) positive examples of gene relationships for the human (resp., mouse) organism;

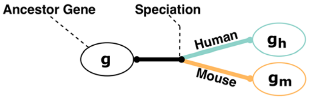

Figure 2. Speciation of an ancestor gene g into two genes g_h and g_m. g_h and g_m are orthologs.

- $orth_{hm}$: $G_h \rightarrow G_m$ (resp., $orth_{mh}$: $G_m \rightarrow G_h$) be a function that takes a human gene $g_h \in G_h$ (resp., a mouse gene $g_m \in G_m$) and returns the corresponding orthologous mouse gene (resp., human gene);
- vec_h: $G_h \rightarrow \mathbb{R}^r$ (resp., vec_m: $G_m \rightarrow \mathbb{R}^q$) be a function that returns the r-dimensional (resp. q-dimensional) vector of expression levels of a human (resp., mouse) gene;
- $^\frown$: $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \rightarrow \mathbb{R}^{n_1+n_2}$ be a function that takes as input two vectors in $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$, respectively, and returns their concatenation in $\mathbb{R}^{n_1+n_2}$.
- $\widetilde{vec}_h$: $G_h \times G_h \rightarrow \mathbb{R}^{2r}$ (resp., $\widetilde{vec}_m$: $G_m \times G_m \rightarrow \mathbb{R}^{2q}$), be a function that takes as input two human (resp. mouse) genes and returns the concatenation of their vectors of expression levels, representing the features of their interaction. Formally, $\widetilde{vec}_h(g'_h, g''_h) = vec_h(g'_h)^\frown vec_h(g''_h)$ and $\widetilde{vec}_m(g'_m, g''_m) = vec_m(g'_m)^\frown vec_m(g''_m)$.

The task solved by our multi-task learning approach is to find two regression functions, namely:

- f_h: $\mathbb{R}^{2r} \rightarrow [0, 1]$, that, given a pair of human genes $g'_h \in G_h$ and $g''_h \in G_h$ represented through the feature vector of their interaction $\widetilde{vec}_h(g'_h, g''_h)$, returns a score representing the degree of certainty of the existence of the interaction between g'_h and g''_h in the human GRN.
- f_m: $\mathbb{R}^{2q} \rightarrow [0, 1]$, that, given a pair of mouse genes $g'_m \in G_m$ and $g''_m \in G_m$ represented through the feature vector of their interaction $\widetilde{vec}_m(g'_m, g''_m)$, returns a score representing the degree of certainty of the existence of the interaction between g'_m and g''_m in the mouse GRN.

Our goal is to learn both predictive functions simultaneously, by considering the degree of certainty of a given gene pair for the human and the mouse organisms as two different target variables of the same training example. It is noteworthy that this choice allows our method to capture possible dependencies that may exist in the output space (i.e., between the target variables). Specifically, we learn a single regression function f_{hm} that takes as input a pair of genes represented according to the features related to both organisms, and returns the degree of certainty for both organisms. Formally:

$$f_{hm} : \mathbb{R}^{2r+2q} \rightarrow [0, 1] \times [0, 1] \tag{1}$$

Note that the construction of *all-in-one* training examples that can be used to learn a multi-target prediction model needs an additional step, i.e., the identification of a match between human genes and mouse genes. In the following subsections, we describe (i) the details of such a step, (ii) the strategy we adopt to solve the issues of the positive-unlabeled setting, (iii) the construction of the dataset used for learning the multi-target regression function f_{hm}, and (iv) the proposed predictive approach.

Orthologous matching and construction of positive training examples. The first step of our method consists of the identification of possible matches between human and mouse genes. This step is necessary in order to represent each gene pair as a single training example, according to the features (i.e., expression levels) measured for both organisms.

To this aim, we exploit the concept of gene orthology. Ortholog genes are genes of different species that are the result of the speciation of the same originating gene (see Fig. 2). Methodologically, we iterate over the human genes $g_m \in G_m$ and identify the corresponding orthologous gene in the mouse organism (Algorithm 1, Lines 2–27).

At the end of this step, we obtain two new sets of genes: $G_{ho} \subseteq G_h$ consisting of the human genes that have orthologs in the set of mouse genes G_m, and $G_{mo} \subseteq G_m$ consisting of mouse genes that have orthologs in the set of human genes G_h.

The subsequent steps of the method work on the orthologous subsets of genes G_{ho} and G_{mo}.

From a machine learning viewpoint, the set of genes corresponds to the set of nodes of the GRNs. However, our unit of analysis is a pair of genes, for which we want to estimate/predict the degree of existence of the interaction. Given that a human gene g_h is described as a vector of expression levels $vec_h(g_h)$, we represent a pair of genes g'_h and g''_h as the concatenation of their feature vectors, i.e., $\widetilde{vec}_h(g'_h, g''_h) = vec_h(g'_h)^\frown vec_h(g'_h)$ (the same

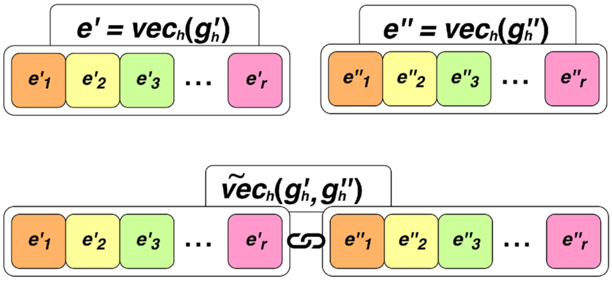
e' = vec_h(g'_h)
e'_1 e'_2 e'_3 ... e'_r
e'' = vec_h(g''_h)
e''_1 e''_2 e''_3 ... e''_r
~vec_h(g'_h, g''_h)
e'_1 e'_2 e'_3 ... e'_r e''_1 e''_2 e''_3 ... e''_r

Figure 3. Concatenation of the expression vectors $e' = vec_h(g'_h)$ and $e'' = vec_h(g''_h)$ of the human genes g'_h and g''_h (the same holds for mouse genes).

holds for the mouse organism). In this step, we build two sets of positive examples P_h and P_m (for human and mouse, respectively) by considering all the pairs of genes appearing in the set of validated interactions B_h and B_m (for human and mouse, respectively), for which we found a matching ortholog in the previous step. We then associate them with the corresponding feature vector (see Fig. 3 and Algorithm 1, Lines 8–12).

Labeling of unlabeled examples.

We recall that we work in the positive-unlabeled setting: together with the (positive) examples identified in the previous step, we also build a set of examples for which we do not have information about the existence of the interaction (unlabeled examples). Following the literature[17], we assign a degree of certainty equal to 1.0 to positive labelled examples, and we estimate the degree of certainty for unlabeled examples according to their similarity with positive examples. Note that, differently from[17], in this work we do not exploit such a similarity to assign a weight to the examples, but to assign a value to their target attributes.

Methodologically, we identify two sets of clusters C_h and C_m, from the human and mouse positive interactions P_h and P_m, respectively, that possibly represent different sub-concepts of existing gene interactions, and exploit them to estimate the value of the target variables of unlabeled examples.

In particular, given two feature vectors u_h (for the human organism) and u_m (for the mouse organism) for the same pair of genes, we compute the value of the target variables t_h and t_m as follows:

$$\begin{aligned} t_h(u_h) &= \max_{c \in C_h} sim(u_h, cent(c)) \\ t_m(u_m) &= \max_{c \in C_m} sim(u_m, cent(c)) \end{aligned} \tag{2}$$

where $cent(c)$ is the feature vector of the centroid of the cluster c, and $sim\colon \mathbb{R}^n \times \mathbb{R}^n \rightarrow [0, 1]$ is a vector similarity function. In this paper we use $sim(a, b) = 1 - \frac{\sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}}{n}$, based on the Euclidean distance, after applying a min-max normalization (in the range [0, 1]) to all the features.

The identification of the clusters can actually be performed through any centroid-based clustering approach. In this paper we rely on the well-known *k-means* algorithm. Moreover, in order to optimally identify the number of clusters k_h for the human organism and k_m for the mouse organism, we use the silhouette cluster analysis[52]. Formally, we define a function $sil\colon P \rightarrow [1, 2, \ldots |P|]$, that, given a set of positive examples $P \in \{P_h, P_m\}$ and the clustering algorithm (in our case, *k-means*) returns the optimal number of clusters, according to the silhouette analysis. In Algorithm 1, this step is performed at Lines 13–14, whereas the exploitation of the identified clusters for computing t_h and t_m is performed at Lines 17–23.

After this step, the main issues of the positive-unlabeled setting are solved, since all the examples are associated to a (known or estimated) value for the target variables t_h and t_m.

Learning the predictive model.

The final stage consists of learning the predictive model, in the form of a multi-target regression function, where the two target variables t_h and t_m represent the degrees of certainty of the existence of the interaction in the human and in the mouse organisms, respectively. With this aim, we build the final training set by concatenating, for each pair of genes for which we identified an ortholog match, (i) the $2r$-dimensional feature vector associated to the human organism, (ii) the $2q$-dimensional feature vector associated to the mouse organism, (iii) the target variable t_h, and (iv) the target variable t_m, leading to training examples represented in $\mathbb{R}^{2r+2q}$, associated to two target variables (see Algorithm 1, Line 24).

Algorithm 1: Our multi-task method for the simultaneous reconstruction of human and mouse GRNs.

```
Data:
    ·$G_h, G_m$: human and mouse gene sets;
    ·$B_h, B_m$: sets of validated human and mouse gene interactions.
Result:
    ·$f_{hm}$: $\mathbb{R}^{2q+2r} \rightarrow [0,1] \times [0,1]$: a predictive function that takes two genes represented according to the features related to both the human and the mouse organisms, and returns the degree of certainty that the link between them is present in the GRNs.
begin
    /* Orthologous matching */
    $G_{ho} \leftarrow \emptyset$; $G_{mo} \leftarrow \emptyset$;
    foreach $g_h \in G_h$ do
        $g_{mo} \leftarrow orth_{hm}(g_h)$;
        if $g_{mo} \neq null$ then
            $G_{ho} \leftarrow G_{ho} \cup \{g_h\}$;
            $G_{mo} \leftarrow G_{mo} \cup \{g_{mo}\}$;

    /* Construct positive examples */
    $P_h \leftarrow \emptyset$;   $P_m \leftarrow \emptyset$;
    foreach $g'_h, g''_h \in B_h \cap (G_{ho} \times G_{ho})$ do
        $P_h \leftarrow P_h \cup \{v\tilde{e}c_h(g'_h, g''_h)\}$;
    foreach $g'_m, g''_m \in B_m \cap (G_{mo} \times G_{mo})$ do
        $P_m \leftarrow P_m \cup \{v\tilde{e}c_m(g'_m, g''_m)\}$;
    /* Identify clusters of positive examples */
    $k_h \leftarrow sil(P_h)$;   $k_m \leftarrow sil(P_m)$;
    $C_h \leftarrow kmeans(P_h, k_h)$; $C_m \leftarrow kmeans(P_m, k_m)$;
    /* Build a multi-target training set */
    $trainingset \leftarrow \emptyset$;
    foreach $\langle g'_{ho}, g''_{ho} \rangle \in G_{ho} \times G_{ho}$ do
        /* Compute the value of the target attribute for the human organism */
        $t_h \leftarrow 1.0$;
        if $v\tilde{e}c_h(g'_{ho}, g''_{ho}) \notin P_h$ then
            $t_h \leftarrow \max_{c \in C_h} sim(v\tilde{e}c_h(g'_{ho}, g''_{ho}), cent(c))$;

        /* Compute the value of the target attribute for the mouse organism */
        $g'_{mo} \leftarrow orth_{hm}(g'_{ho})$;   $g''_{mo} \leftarrow orth_{hm}(g''_{ho})$;
        $t_m \leftarrow 1.0$;
        if $v\tilde{e}c_m(g'_{mo}, g''_{mo}) \notin P_m$ then
            $t_m \leftarrow \max_{c \in C_m} sim(v\tilde{e}c_m(g'_{mo}, g''_{mo}), cent(c))$;
        /* Construction of the multi-target example */
        $e \leftarrow \langle v\tilde{e}c_h(g'_{ho}, g''_{ho})^\frown v\tilde{e}c_m(g'_{mo}, g''_{mo}), t_h, t_m \rangle$;
        $trainingset \leftarrow trainingset \cup \{e\}$;
    /* Train and return the multi-target model */
    $f_{hm} \leftarrow PCT(trainingset)$;
    return $f_{hm}$
```

We learn the predictive model with CLUS[50], that is based on Predictive Clustering Trees (PCTs). We induce PCTs through a standard approach for the top-down induction of decision/regression trees, that takes as input a set of training examples and returns the induced tree. The heuristics adopted to select the best tests of the internal nodes of the tree is the reduction of variance achieved by partitioning the examples according to such a test. The maximization of the variance reduction leads to maximizing the cluster homogeneity and, therefore, to improving the predictive performance. Therefore, the considered heuristic is formally defined as $Var_E(t_h, t_m) = Var_E(t_h) + Var_E(t_m)$,, where $Var_E(t_h)$ (resp., $Var_E(t_m)$) is the variance observed on the target attribute t_h (resp., t_m) on the set of examples E falling in a given node of the tree.

This means that the variance reduction, used to identify the best candidate split in the tree construction, is computed as:

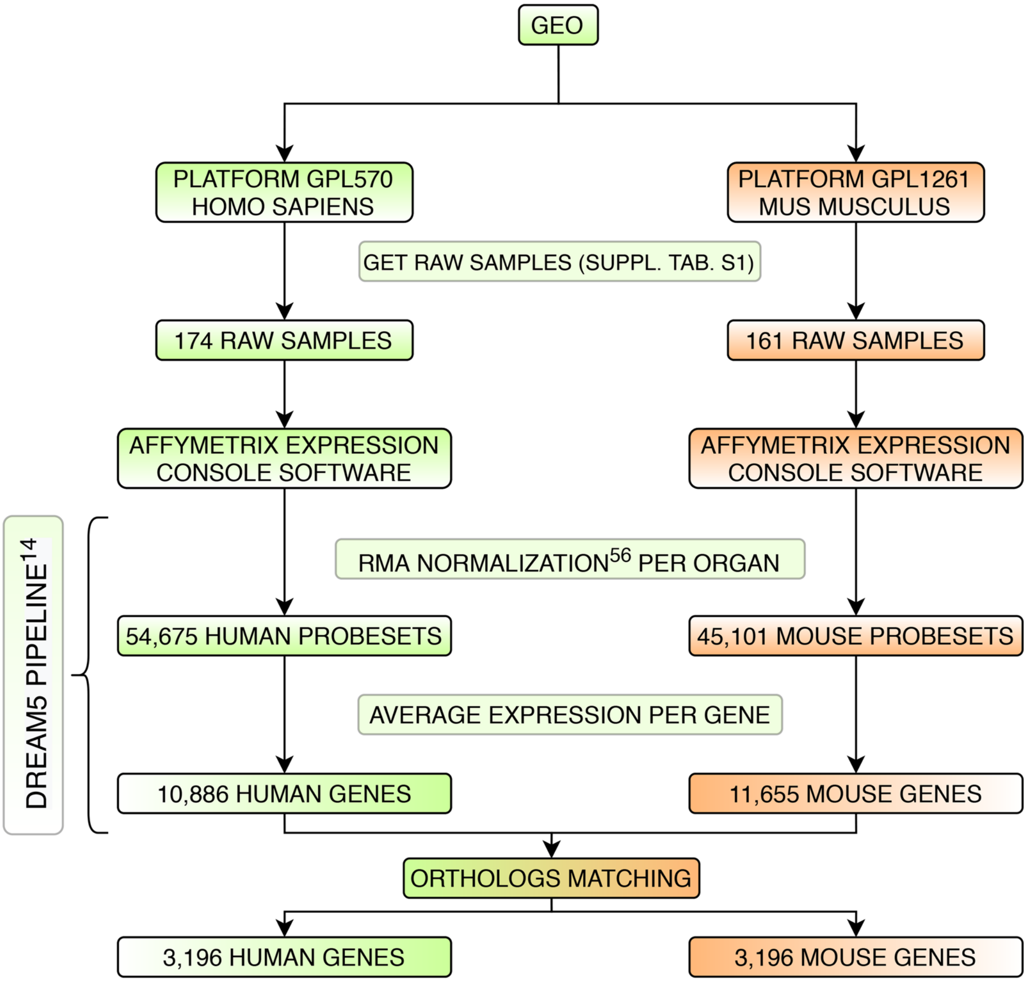
GEO
PLATFORM GPL570 HOMO SAPIENS
PLATFORM GPL1261 MUS MUSCULUS
GET RAW SAMPLES (SUPPL. TAB. S1)
174 RAW SAMPLES
161 RAW SAMPLES
AFFYMETRIX EXPRESSION CONSOLE SOFTWARE
AFFYMETRIX EXPRESSION CONSOLE SOFTWARE
DREAM5 PIPELINE14
RMA NORMALIZATION56 PER ORGAN
54,675 HUMAN PROBESETS
45,101 MOUSE PROBESETS
AVERAGE EXPRESSION PER GENE
10,886 HUMAN GENES
11,655 MOUSE GENES
ORTHOLOGS MATCHING
3,196 HUMAN GENES
3,196 MOUSE GENES

Figure 4. Graphical overview of the pipeline followed to identify the genesets and their gene expression levels.

	Human	Mouse
Probeset	54,675	45,101
Geneset	10,886	11,655
Orthologous geneset	3,196	3,196
Gene features	174	161
Gene-pair features	348	322
Positive examples	3,970	3,970
Unlabeled examples	75,430	75,430

Table 1. Quantitative characteristics of the dataset.

Dataset	50%	40%	30%	20%	10%	5%
Positive	3,970	3,970	3,970	3,970	3,970	3,970
Unlabeled	3,970	5,558	9,263	15,880	35,730	75,430

Table 2. Number of examples for each variant of the dataset.

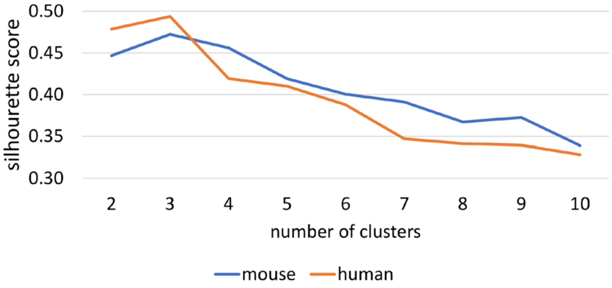

Figure 5. Silhouette score for the human (k_h) and mouse (k_m) organisms.

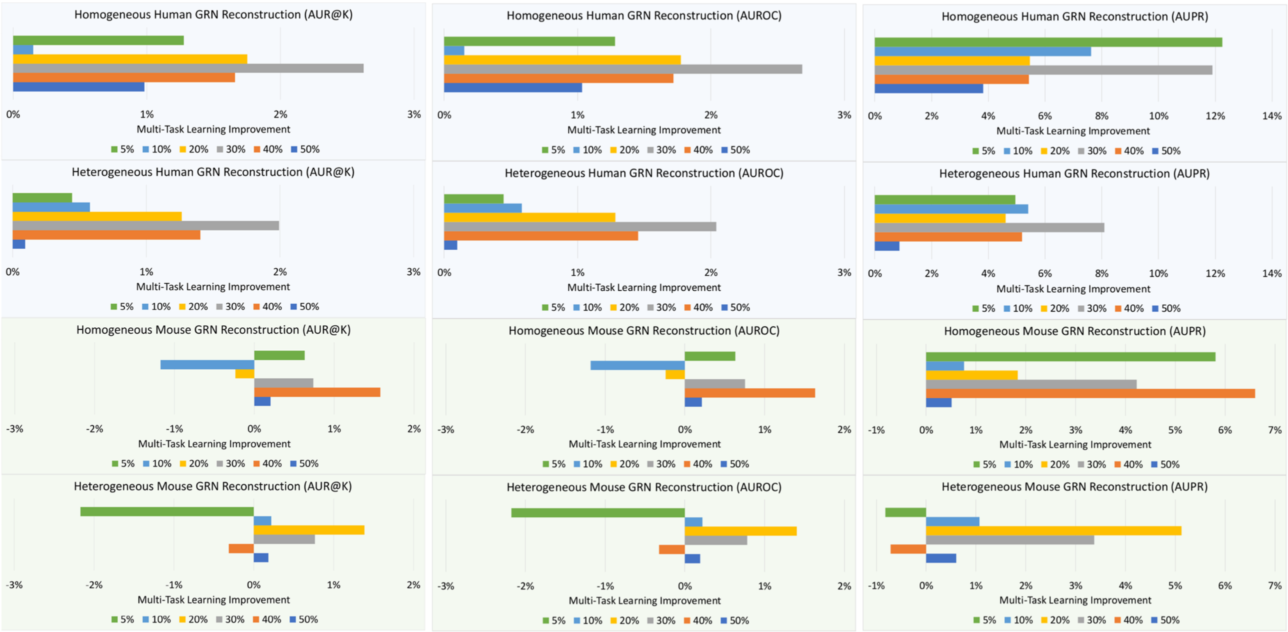

Figure 6. Improvement achieved by our approach with respect to no_transfer, in terms of AUR@K, AUROC and AUPR.

$$h = Var_E(t_h, t_m) - (Var_{E'}(t_h, t_m) + Var_{E''}(t_h, t_m)) \tag{3}$$

where E, E', E'' are the sets of examples in the parent, left and right child nodes, respectively.

The final learned model will represent the function f_{hm}: $\mathbb{R}^{2r+2q} \rightarrow [0,1] \times [0,1]$ that can be employed to predict the degree of certainty for all the unlabeled examples for both considered organisms. Therefore, as explained in "Methods" section, f_{hm} acts like the two functions f_h and f_m simultaneously, with the advantage of being able to capture dependencies in the output space.

Experiments

In this section, we present the results of our experimental evaluation. All the experiments were performed on a server equipped with a 6-cores CPU @ 3.50Ghz and 128GB RAM. In the following subsections, we first describe the considered competitor systems, the datasets and the experimental setting. Finally, we present and discuss the obtained results.

Competitor approaches.

We compared our method with the following competitor approaches:

- **TJM**[29], that reduces the difference between the two domains by identifying a match between their features and by reweighting the instances to construct a new reduced/shared feature representation;
- **BDA**[31], that adaptively leverages the importance of the marginal and conditional distribution discrepancies between the two domains;
- **JGSA**[30], that learns two coupled projections, that are exploited to project the source and the target domain data into low-dimensional subspaces, where the geometrical and the distribution discrepancies are minimized;

Measure	50%	40%	30%	20%	10%	5%
Homogeneous human GNR						
AUROC	V	V	V	V	V	V
AUPR	V	V	V	V	V	V
AUR@K	V	V	V	V	V	V
Heterogeneous human GNR						
AUROC	V	V	V	V	V	V
AUPR	V	V	V	V	V	V
AUR@K	V	V	V	V	V	V
Homogeneous mouse GNR						
AUROC	V	V	V	V	X	V
AUPR	V	V	V	V	V	V
AUR@K	V	V	V	X	X	V
Heterogeneous mouse GNR						
AUROC	V	X	V	V	V	X
AUPR	V	X	V	V	V	X
AUR@K	V	X	V	V	V	X

Table 3. Summary of settings for which the multi-task approach provided an improvement over the baseline.

- **no_transfer**, that is the single-domain variant of our approach, which reconstructs each single GRN independently.

TJM, BDA, and JGSA are feature-based transfer learning methods that are able to share the knowledge between (also) heterogeneous source and target domains, as long as they are described with the same number of features. On the other hand, the no_transfer approach can be considered a baseline, that allows us to evaluate the positive contribution of the multi-target approach proposed in this paper or, conversely, to evaluate the possible presence of *negative transfer* phenomena[53,54], where the use of knowledge coming from other domains actually compromises the quality of the reconstruction.

Datasets. We built the dataset by downloading a compendium of microarray data of both human (Platform ID: GPL570) and mouse (Platform ID: GPL1261) organisms from Gene Expression Omnibus—GEO (www.ncbi.nlm.nih.gov/geo/), a publicly available web repository hosted by the National Center for Biotechnology Information (NCBI). In total, 174 and 161 raw CEL files related to 54,675 and 45,101 control probesets of 6 different organs were downloaded for human and mouse organisms, respectively (see Supplementary Table S1 for a complete list of accession numbers). More specifically, the 174 CEL files for the human organism are distributed as follows: 17 for bone marrow, 37 for brain, 4 for heart, 7 for liver, 45 for lung, and 64 for skin. On the other hand, the 161 CEL files of the mouse organism are distributed as follows: 14 for bone marrow, 8 for brain, 8 for heart, 124 for liver, 4 for lung, and 3 for skin.

We processed the data following the workflow proposed in the DREAM5 challenge[14] (see Fig. 4 for a graphical overview of the followed pipeline). In particular, we performed the Robust Multi-array Average (RMA)[55] normalization through Affymetrix Expression Console Software as one batch per organ. Data were background adjusted, quantile normalized, median polished and log-transformed. The mapping from Affymetrix probeset IDs to gene IDs led to a total of 10,886 human genes and to 11,655 mouse genes.

After the ortholog matching (see "Orthologous Matching and construction of positive training examples" section), we obtained a reduced set of 3, 196 genes for both organisms. The strategy adopted to perform the ortholog matching was based on the gene symbol, that corresponds between homo sapiens and mus musculus organisms, except for differences in the capitalization[56]. Alternative solutions may have been adopted, mostly based on explicit lists of ortholog genes (see, for example, the OMA orthology database[57]), but the strategy based on the gene symbol provided us 153 additional matches. Note that the set of 3,043 genes was included in the set of 3,196 genes we considered.

The dataset of possible relationships between genes was built by considering all the possible pairs of genes (more than 10 million, excluding the self-links), each associated with the concatenation of the feature vectors of the involved genes (following the strategy described in "Methods" section). This step led to 348-dimensional vectors for human gene pairs and to 322-dimensional vectors for mouse gene pairs. We exploited the database BioGRID[20] as the source of known validated interactions (i.e., to define the sets B_h and B_m), while the remaining possible relationships were considered unlabeled. In Table 1 we report a summary of the quantitative characteristics of the considered dataset.

Since some competitor systems, even if they are able to work in the heterogeneous transfer learning setting, require the number of features of all the domains to correspond, we also built a homogeneous version of the dataset. In particular, we aggregated the features associated to each organ (by averaging their value), leading to a homogeneous dataset consisting of 6 features per gene, for both the human and the mouse organisms.

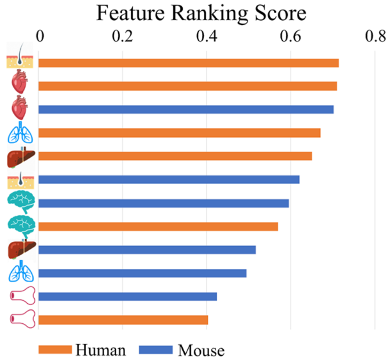

Figure 7. Ranking of the features in the homogeneous setting.

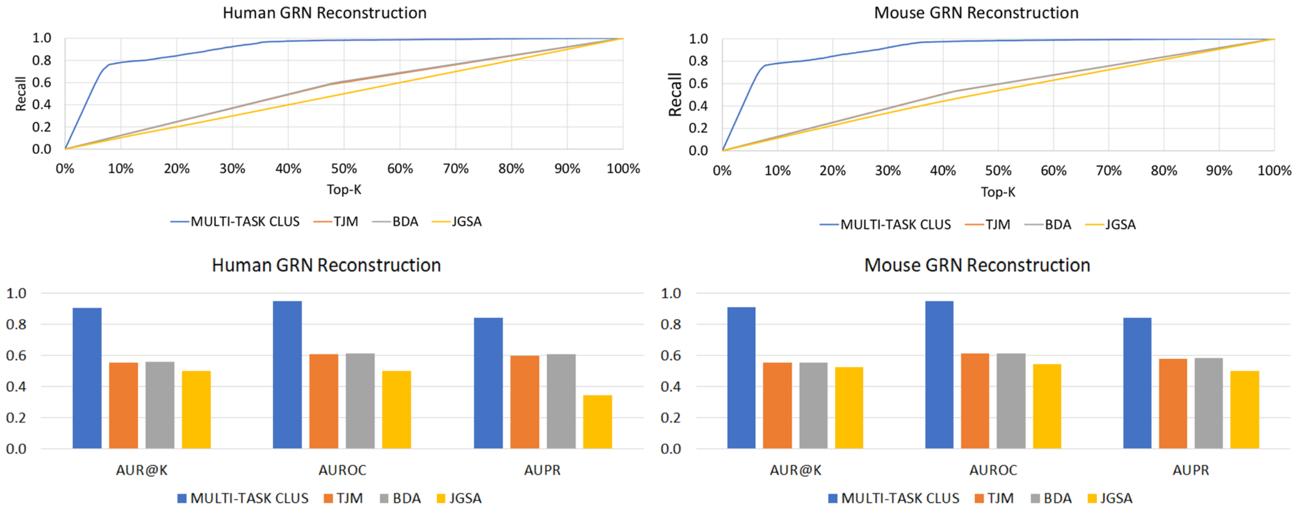

Figure 8. Recall@k (charts in the top) and AUR@K, AUROC and AUPR (charts in the bottom) results, obtained by our method (referred to as MULTI-TASK CLUS here) and its competitors.

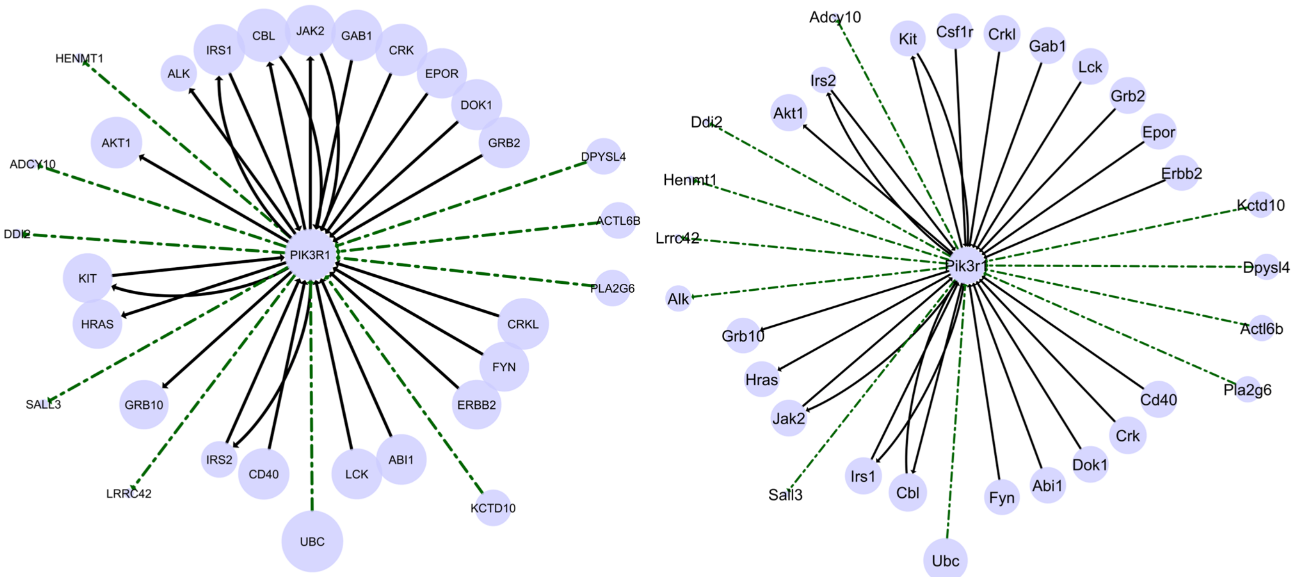

Figure 9. The subnetwork identified for the human hub gene PIK3R1 (on the left) and the mouse hub gene Pik3r1 (on the right). The size of the circle of a gene represents the number of genes regulated by such a gene. Known interactions (from BioGRID) are represented as black lines, while interactions predicted by our system are represented as green dotted lines.

Finally, we evaluated the robustness of the proposed method with respect to the number of unlabeled examples used during the training phase. With this aim, we considered a sample of maximum 75, 430 unlabeled examples and we built 6 variants of the datasets, with a different ratio of the number of positive examples over the unlabeled examples (see Table 2). It is worth mentioning that the comparison with competitors has been performed on the smallest version of the dataset (i.e., with the ratio 50%), because they were not able to complete the experiments with larger datasets without incurring in RAM exhausting errors on our servers. Indeed, while our approach is based on a top-down induction of regression trees, that is generally efficient, competitor methods are mainly based on matrix computations and easily exhausted the RAM on our servers even with the considered reduced dataset.

Experimental setting. The results have been produced according to a 10-fold cross-validation approach, where each fold consists of 9/10 of the positive examples for training and 1/10 of the positive examples for testing, while all the unlabeled examples are considered for both training and testing. The preliminary estimation of the optimal value of k_h and k_m for our method led to the results reported in Fig. 5. Accordingly, we considered the best configurations $k_h = 3$ and $k_m = 3$ for the subsequent experiments.

Since we work in the positive-unlabeled learning setting where no negative examples are available[16,17,58], we evaluate the performance of different methods in terms of recall@k and the area under the recall@k curve (AUR@K). The recall@k measure is defined as $\frac{TP_k}{TP+FN}$, where TP_k is the number of returned true positive interactions, within the first top-k interactions, and $(TP + FN)$ corresponds to the number of positive examples in the testing fold. This formula allows us to evaluate the ability of the method to put reliable interactions on the top part of the returned ranking. The recall@k curve is a curve representing the recall@k with varying values of k, while the AUR@K measure corresponds to the area under such a curve.

We also report the results in terms of the area under the ROC curve (AUROC) and the area under the precision-recall curve (AUPR). Note that the recall@k and the AUR@K measures do not introduce any (possibly wrong) bias in the ground truth, while in the computation of the AUROC and AUPR measures, it is necessary to assume the unlabeled examples as negative examples.

Results. Figure 6 depicts the improvement obtained by our approach with respect to the baseline no_transfer, in terms of the AUR@K, AUROC and AUPR measures, for both the homogeneous and heterogeneous datasets, and with respect to both organisms.

The charts show that the proposed approach provides a marked improvement over the single-domain counterpart in the reconstruction of the human GRN. Such an advantage is evident for all the variants of the dataset, i.e., for all the considered labeled/unlabeled ratios. On the other hand, the reconstruction of the mouse GRN appears to exploit the knowledge about the human GRN only with higher labeled/unlabeled ratios. This may suggest that the mouse organism can be considered an appropriate model organism for the study of the human GRN, but the contrary may hold to a lesser degree, i.e., only when there is a sufficient amount of biologically validated information.

In general, for both organisms, the higher the labeled/unlabeled ratio, the better the quality of the reconstruction. This is an expected result, since the unlabeled examples could belong to clusters of positive examples that were not properly represented/observed in the set of positive examples, due to their limited availability. Despite such an expected result, the reconstruction performed with our multi-task approach in most of the cases provides an advantage, even with very low labeled/unlabeled ratios. The few cases in which there is not an improvement occur in the reconstruction of the mouse regulatory network. This may indicate that it is easier and more natural to exploit the knowledge coming from a general/simple organism to describe a complex organism, rather than vice versa. In Table 3 we show a summary of the settings where our multi-task approach provided an improvement over the baseline.

Moreover, as explained in "Competitor approaches" section, we compared our results with those obtained by three state-of-the-art approaches. In Fig. 8 we present the results in terms of recall@k, AUR@K, AUROC and AUPR. Observing the figure, it is clear that our approach outperforms all the competitors by a large margin for both the considered organisms.

We also performed a qualitative evaluation of the networks reconstructed by our system. For this specific analysis, we considered the largest version of the dataset, containing 3, 970 positive examples and 75, 430 unlabelled examples (see the variant 5% in Table 2). Since the experiments were performed using 10-fold cross validation, resulting in 10 different rankings (one for each fold), we averaged the scores and analyzed the resulting ranking. We then selected the top 10, 000 ranked interactions for both organisms, we computed some topological measures (see Supplementary Tables S2 and S3 for a detailed overview) and we identified the hub genes, by selecting the top-10% of genes with the highest numbers of regulated genes[59]. Among them, we selected the 352 genes appearing as hubs for both the considered organisms, and we plotted the subnetworks involving each of them, emphasizing the interactions that were present in BioGRID (in black) and those that were predicted by our system (in green). It is noteworthy that the subnetworks of the first 60 hub genes are identical, in both known and predicted interactions, between the two organisms. This confirms the ability of our system in catching cross-organism similarities and in predicting the existence of interactions that appear coherent among the organisms.

In Fig. 9 we depict the first subnetwork that shows some differences between the organisms (i.e., the 61th in the ranking). This is the case of the human gene PIK3R1 (resp., Pik3r1 for the mouse organism). In this case, we can observe 5 (resp., 6) predicted regulated genes for the human (resp., for the mouse) organism and 5 (resp., 5) predicted genes regulating PIK3R1 (resp., Pik3r1). Specifically, it is noteworthy that the interaction Pik3r1 → Alk has been inferred by our method, but is not covered in BioGRID. On the other hand, the interaction Csf1r → Pik3r1 is present in BioGRID for the mouse organism, but our method did not suggest the corresponding interaction for the human organism, that is actually absent in BioGRID (preventing, therefore, a possible false positive). This confirms that, although our method exploits the knowledge coming from the simultaneous reconstruction of the regulatory networks of both the organisms, it does not merely mimic the behavior observed on an organism on the other one. On the contrary, it is able to catch possible differences and asymmetries.

Finally, we performed an additional analysis regarding the importance of the considered features. In particular, focusing on the homogeneous dataset, we aimed to identify the most relevant organs (i.e., those most relevant to our method during the learning of the multi-target regression model), following the approach of Petković et al.[60]. In Fig. 7 we show the obtained ranking, where we can observe that the features associated to the human skin and heart, together with those associated to the mouse heart and lung, have been detected as the most relevant ones for the gene network reconstruction task. In contrast, it seems that features related to bone marrow (for both organisms) did not provide any relevant contribution. In the middle, we find the features closely related to the brain (for both organisms), the human liver and the human lungs. This behaviour is probably motivated by the fact that some organs show more similar properties between the two organisms, or are better connected through orthologous genes, than others. It is noteworthy that these findings can be profitably exploited to focus future work, where larger sets of samples, related to the organs that provide a higher contribution, can be adopted.

Conclusion

Several computational approaches, mainly based on machine learning methods, can be employed for the reconstruction of GRNs. However, existing gene network reconstruction methods suffer when the number of labeled examples is low, especially when no negative examples are available. In this paper we have proposed a method that overcomes these limitations. Our approach is able to simultaneously reconstruct the GRN of two organisms, by exploiting a multi-target regression approach that, in conjunction with the concept of gene orthology, is able to natively work in a positive-unlabeled learning setting.

The experiments show that our approach is able to really "transfer" knowledge extracted from an organism and profitably use it in another organism. Moreover, the proposed multi-target positive-unlabeled learning algorithm outperforms both its single-task counterpart and three state-of-the-art transfer learning approaches in the reconstruction of both GRNs.

As future work we plan to define a more general approach to map the examples in the considered domains, so that it may be adopted in multiple, even non-biological, applications. Moreover, while in the present paper we presented our novel approach and evaluated its effectiveness, compared with state-of-the-art methods, we plan to extend the experiments to larger datasets, also considering different pipelines.

Received: 10 April 2020; Accepted: 29 October 2020
Published online: 18 December 2020

References

1. Sager, R. Expression genetics in cancer: Shifting the focus from DNA to RNA. *Proc. Nat. Acad. Sci.* **94**, 952–955 (1997).
2. Ross, D. T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* **24**, 227–235 (2000).
3. Davidson, E. H. & Peter, I. S. Gene regulatory networks. In *Genomic Control Process* 41–77 (Elsevier, Amsterdam, 2015).
4. Smith, V. A., Jarvis, E. D. & Hartemink, A. J. Evaluating functional network inference using simulations of complex biological systems. *Bioinformatics* **18**, S216–S224 (2002).
5. Stolovitzky, G., Monroe, D. & Califano, A. Dialogue on reverse-engineering assessment and methods. *Ann. N. Y. Acad. Sci.* **1115**, 1–22 (2007).
6. Park, P. J. Chip-seq: Advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **2**, 669–680 (2009).
7. Bulyk, M. L. Discovering DNA regulatory elements with bacteria. *Nat. Biotechnol.* **23**, 942–944 (2005).
8. Berger, M. F. & Bulyk, M. L. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc.* **4**, 393–411 (2009).
9. Penfold, C. A. & Wild, D. L. How to infer gene networks from expression profiles, revisited. *Interface Focus* **1**, 857–870 (2011).
10. Emmert-Streib, F. *et al.* Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Bioinform. Comput. Biol.* **3**, 8 (2012).
11. Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E. & Guthke, R. Gene regulatory network inference: data integration in dynamic models—a review. *Biosystems* **96**, 86–103 (2009).
12. Markowetz, F. & Spang, R. Inferring cellular networks—a review. *BMC Bioinform.* **8**, 2 (2007).
13. De Jong, H. Modeling and simulation of genetic regulatory systems: A literature review. *J. Comput. Biol.* **9**, 67–103 (2002).
14. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 (2012).
15. Ceci, M., Pio, G., Kuzmanovski, V. & Džeroski, S. Semi-supervised multi-view learning for gene network reconstruction. *PLoS One* **10**, 1–27 (2015).
16. Mignone, P. & Pio, G. Positive unlabeled link prediction via transfer learning for gene network reconstruction. *ISMIS* **2018**, 13–23 (2018).
17. Mignone, P., Pio, G., Delia, D. & Ceci, M. Exploiting transfer learning for the reconstruction of the human gene regulatory network. *Bioinformatics* **36**, 1553–1561 (2020).
18. Weiss, K. R., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *J. Big Data* **3**, 9 (2016).
19. Koonin, E. Orthologs, paralogs, and evolutionary genomics 1. *Annu. Rev. Genet.* **39**, 309–38 (2005).
20. Stark, C. *et al.* Biogrid: A general repository for interaction datasets. *Nucleic Acids Res.* **34**, 535–539 (2006).
21. Zhang, B. & Zuo, W. Learning from positive and unlabeled examples: A survey. In *2008 International Symposiums on Information Processing*, 650–654 (2008).
22. Liu, B., Dai, Y., Li, X., Lee, W. S. & Yu, P. S. Building text classifiers using positive and unlabeled examples. *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), 19-22 December 2003, Melbourne, Florida, USA* 179–188 (2003).
23. Yu, H., Han, J. & Chang, K.-C. Pebl: Positive example based learning for web page classification using svm. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 239–248 (2002).
24. Li, X. & Liu, B. Learning to classify texts using positive and unlabeled data. *IJCAI International Joint Conference on Artificial Intelligence***587–592**, (2003).

25. Elkan, C. & Noto, K. Learning classifiers from only positive and unlabeled data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 213–220 (2008). Cited By :323.
26. Lee, W. S. & Liu, B. Learning with positive and unlabeled examples using weighted logistic regression. In *Proceedings, Twentieth International Conference on Machine Learning*, vol. 1, 448–455 (2003). Cited By :157.
27. Pio, G., Malerba, D., Delia, D. & Ceci, M. Integrating microrna target predictions for the discovery of gene regulatory networks: A semi-supervised ensemble learning approach. *BMC Bioinform.* **15**, S4 (2014).
28. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
29. Long, M., Wang, J., Ding, G., Sun, J. & Yu, P. S. Transfer joint matching for unsupervised domain adaptation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014*, 1410–1417 (2014).
30. Zhang, J., Li, W. & Ogunbona, P. Joint geometrical and statistical alignment for visual domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 5150–5158 (2017).
31. Wang, J., Chen, Y., Hao, S., Feng, W. & Shen, Z. Balanced distribution adaptation for transfer learning. In *2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017*, 1129–1134 (2017).
32. Huang, J., Smola, A. J., Gretton, A., Borgwardt, K. M. & Schölkopf, B. Correcting sample selection bias by unlabeled data. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*, 601–608 (2006).
33. Jiang, J. & Zhai, C. Instance weighting for domain adaptation in NLP. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (2007).
34. Gao, J., Fan, W., Jiang, J. & Han, J. Knowledge transfer via multiple model local structure mapping. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 283–291 (2008).
35. Bonilla, E. V., Chai, K. M. A. & Williams, C. K. I. Multi-task gaussian process prediction. In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems*, 153–160 (2007).
36. Pan, S. J., Kwok, J. T. & Yang, Q. Transfer learning via dimensionality reduction. *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI* **2008**, 677–682 (2008).
37. Blitzer, J., McDonald, R. T. & Pereira, F. Domain adaptation with structural correspondence learning. In *EMNLP 2006, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 120–128 (2006).
38. Achanta, H. K., Misganaw, B. & Vidyasagar, M. A transfer learning approach for integrating biological data across platforms. In *2016 American Control Conference (ACC)*, 6695–6697 (2016).
39. Sevakula, R. K., Singh, V., Verma, N. K., Kumar, C. & Cui, Y. Transfer learning for molecular cancer classification using deep neural networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics***1–1**, (2018).
40. Breckels, L. M. *et al.* Learning from heterogeneous data sources: An application in spatial proteomics. *PLoS Comput. Biol.* **12**, 1–26 (2016).
41. Collobert, R. & Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, 160–167 (ACM, New York, NY, USA, 2008).
42. Deng, L., Hinton, G. & Kingsbury, B. New types of deep neural network learning for speech recognition and related applications: an overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8599–8603 (2013).
43. Misra, I., Shrivastava, A., Gupta, A. & Hebert, M. Cross-stitch networks for multi-task learning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3994–4003 (2016).
44. Castro, D. M., de Veaux, N. R., Miraldi, E. R. & Bonneau, R. Multi-study inference of regulatory networks for more accurate models of gene regulation. *PLoS Comput. Biol.* **15**, 1–22 (2019).
45. Kaji, H., Yamaguchi, H. & Sugiyama, M. Multi task learning with positive and unlabeled data and its application to mental state prediction. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018*, 2301–2305 (2018).
46. Rei, M. Semi-supervised multitask learning for sequence labeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2121–2130 (Association for Computational Linguistics, Vancouver, Canada, 2017).
47. Levatic, J., Kocev, D., Ceci, M. & Džeroski, S. Semi-supervised trees for multi-target regression. *Inf. Sci.* **450**, 109–127 (2018).
48. Bakir, G. H. *et al. Predicting Structured Data (Neural Information Processing)* (The MIT Press, Cambridge, 2007).
49. Stojanova, D., Ceci, M., Appice, A., Malerba, D. & Džeroski, S. Dealing with spatial autocorrelation when learning predictive clustering trees. *Ecol. Inf.* **13**, 22–39 (2013).
50. Blockeel, H., Raedt, L. D. & Ramon, J. Top-down induction of clustering trees. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998)*, 55–63 (1998).
51. Kocev, D., Vens, C., Struyf, J. & Džeroski, S. Tree ensembles for predicting structured outputs. *Pattern Recogn.* **46**, 817–833 (2013).
52. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
53. Ge, L., Gao, J., Ngo, H. Q., Li, K. & Zhang, A. On handling negative transfer and imbalanced distributions in multiple source transfer learning. *Stat. Anal. Data Min.* **7**, 254–271 (2014).
54. Seah, C., Ong, Y. & Tsang, I. W. Combating negative transfer from predictive distribution differences. *IEEE Trans. Cybern.* **43**, 1153–1165 (2013).
55. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
56. Wright, M. W. & Bruford, E. A. Human and orthologous gene nomenclature. *Gene* **369**, 1–6 (2006).
57. Altenhoff, A. M. *et al.* The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.* **46**, D477–D485 (2017).
58. Pio, G., Ceci, M., Malerba, D. & Delia, D. ComiRNet: A web-based system for the analysis of miRNA-gene regulatory networks. *BMC Bioinform.* **16**, S7 (2015).
59. Liu, Y. *et al.* Identification of hub genes and key pathways associated with bipolar disorder based on weighted gene co-expression network analysis. *Front. Physiol.* **10**, 1081 (2019).
60. Petković, M., Džeroski, S. & Kocev, D. Feature ranking for multi-target regression with tree ensemble methods. In Yamamoto, A., Kida, T., Uno, T. & Kuboyama, T. (eds.) *Discovery Science*, 171–185 (Springer International Publishing, Cham, 2017).

Acknowledgements

This work was supported by the Ministry of Education, Universities and Research through the PON project TALIsMAn - Tecnologie di Assistenza personALizzata per il Miglioramento della quAlitá della vitA (Grant N. ARS01_01116), and by the Slovenian Research Agency through the grants P2-0103 and N2-0128. Dr. Gianvito Pio acknowledges the support of Ministry of Education, Universities and Research (MIUR) through the project "Big Data Analytics", AIM 1852414, activity 1, line 1. Finally, we would like to thank Lynn Rudd for her help in reading the manuscript.

Author contributions

All the authors collaborated to conceive the task and design the solution from a methodological point of view. P.M. implemented the system, ran the experiments and collected the results. G.P., M.C. and S.D. analyzed and discussed the results. M.C. supervised the research activities. All the authors contributed to the manuscript drafting and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-78033-7.

Correspondence and requests for materials should be addressed to G.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.