Data and Text Mining

Integrating genome-scale metabolic modelling and transfer learning for human gene regulatory network reconstruction

Gianvito Pio [1,2,*], Paolo Mignone [1,2], Giuseppe Magazzù [3], Guido Zampieri [3,4], Michelangelo Ceci [1,2,5] and Claudio Angione [3,6,7]

[1] Department of Computer Science, University of Bari Aldo Moro, Via Orabona, 4, 70125, Bari, Italy
[2] Big Data Lab, National Interuniversity Consortium for Informatics (CINI), Rome
[3] School of Computing, Engineering & Digital Technologies, Teesside University, Borough road, TS1 3BA, Tees Valley, UK
[4] Department of Biology, University of Padova, Via U. Bassi 58/b, 35121, Padova, Italy
[5] Department of Knowledge Technologies, Jozef Stefan Institute, Jamova 39, 1000, Ljubljana, Slovenia
[6] Centre for Digital Innovation, Teesside University, Campus Heart, TS1 3BX, Tees Valley, UK
[7] Healthcare Innovation Centre, Teesside University, Campus Heart, TS1 3BX, Tees Valley, UK

[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Gene regulation is responsible for controlling numerous physiological functions and dynamically responding to environmental fluctuations. Reconstructing the human network of gene regulatory interactions is thus paramount to understanding the cell functional organisation across cell types, as well as to elucidating pathogenic processes and identifying molecular drug targets. Although significant effort has been devoted towards this direction, existing computational methods mainly rely on gene expression levels, possibly ignoring the information conveyed by mechanistic biochemical knowledge. Moreover, except for a few recent attempts, most of the existing approaches only consider the information of the organism under analysis, without exploiting the information of related model organisms.

Results: We propose a novel method for the reconstruction of the human gene regulatory network, based on a transfer learning strategy that synergically exploits information from human and mouse, conveyed by gene-related metabolic features generated in-silico from gene expression data. Specifically, we learn a predictive model from metabolic activity inferred via tissue-specific metabolic modelling of artificial gene knockouts. Our experiments show that the combination of our transfer learning approach with the constructed metabolic features provides a significant advantage in terms of reconstruction accuracy, as well as additional clues on the contribution of each constructed metabolic feature.

Availability: The system, the datasets and all the results obtained in this study are available at: `https://doi.org/10.6084/m9.figshare.c.5237687`

Contact: gianvito.pio@uniba.it

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

1 Introduction

Living organisms need, for their survival and replication, a gene regulatory system responsible for their maintenance, development and response to changing environmental conditions. Gene regulation is orchestrated by large sets of regulator molecules with specific targets, which collectively form a gene regulatory network (GRN). The mapping of GRNs was recently propelled by the surge of high-throughput data, that led to both the discovery of unknown biological interactions and a deeper understanding of known structures (Davidson, 2010; Ye *et al.*, 2018; Gardner *et al.*, 2003). The problem of GRN reconstruction has therefore wide applications in basic biology, but also in related disciplines, such as biomedicine and biotechnology (Karlebach and Shamir, 2008).

Several computational methods for GRN reconstruction have been proposed in the literature, including graphical Gaussian models (Schäfer and Strimmer, 2005), Bayesian networks (Zou and Conzen, 2005), as well as approaches that consider and exploit causality phenomena (Pio *et al.*, 2020; Luo *et al.*, 2009) or knowledge derived from related organisms (Mignone *et al.*, 2020a). To predict unknown relationships, GRNs have also been mathematically integrated with metabolic networks, which mediate interactions between gene regulation and environmental cues (Yeang and Vingron, 2006).

At the same time, the importance of GRNs is connected to the long-standing problem of understanding the relationship between genotype and phenotype. Solving this problem can shed light on many open biological questions, such as the etiology of a disease and its potential treatments, or the mechanisms that regulate cell physiology. The most adopted approaches are genome-wide association studies (GWAS) and systems biology methods. GWAS can associate gene variations to phenotypic traits, but fail to provide a mechanistic explanation for their findings (Welter *et al.*, 2014). Systems biology techniques are designed to address this issue (Yurkovich and Palsson, 2015). Widely used systems biology methods are genome-scale metabolic models (GSMMs), i.e., mathematical representations of known biochemical reactions and transmembrane transporters in an organism. Being focused on metabolism, they biologically complement GRN models, and are indeed suited for integration (Chandrasekaran and Price, 2010; Motamedian *et al.*, 2017b).

A peculiarity of GSMMs is that they allow capturing long-range phenomena on the scale of cellular systems thanks to the functional information they contain, encoded in their metabolic pathway and reaction representations (Richelle *et al.*, 2020). Flux balance analysis (FBA) is one of the most employed techniques to estimate the metabolic activity within a GSMM at steady-state (i.e., when metabolite concentrations do not change). This enables the conversion of the mathematical approach into a linear system, which is solved for the reaction rates, namely the activity of each reaction in the network (Palsson, 2015). The additional advantage of this approach is that GSMMs can be tailored to a specific condition or individual. This is achieved by constraining the GSMM using experimental data and omics profiles, thereby creating a context-specific model from which to draw conclusions for a specific experimental setting (Töpfer *et al.*, 2015; Vijayakumar *et al.*, 2018; Nielsen, 2017). These models, exploited by methods such as FBA, can provide context-specific metabolic reaction fluxes, whose information can be used coupled with other relevant omic data to make predictions or improve the understanding of several biological phenomena (Magazzù *et al.*, 2021; Yang *et al.*, 2019).

There is a growing trend in the adoption of machine learning methods in biology Tonkovic *et al.* (2020) and, specifically, to process and interpret the output of metabolic models (Zampieri *et al.*, 2019; Kavvas *et al.*, 2020; Yang *et al.*, 2019; Culley *et al.*, 2020; Ben Guebila and Thiele, 2019). *In silico* metabolic information has been adopted in reconstructing GRNs (Karlebach and Shamir, 2008; Schlitt and Brazma, 2007), and to infer gene relationships by analysing the metabolic effect of simultaneous gene KO (Wang *et al.*, 2017; Occhipinti *et al.*, 2020). However, metabolic network modelling has not been used to inform GRN inference methods in combination with transfer learning.

Following this line of research, in this paper we investigate the potential of the exploitation of metabolic information while reconstructing the human GRN, in an integrated transfer learning framework. In particular, we reconstruct the human GRN by leveraging the knowledge about an additional model organism (Mignone *et al.*, 2020b), i.e., the mouse, and exploit both a set of known/verified regulations as well as a large set of still unstudied gene regulations. The two considered organisms are linked by considering their orthologous genes, i.e., genes inherited in both species from a common ancestor gene. Such genes are integrated within a constraint-based model that simulates their artificial knockout and determines how this perturbation propagates over the corresponding metabolic network. This approach allows us to catch possible analogies between the organisms in terms of their metabolic fluxes, in both known and still unknown regulations. Our experimental evaluation, described in detail in Section 3, empirically proves the effectiveness of the proposed integrated approach, in terms of both increased accuracy of the reconstruction and of possible clues coming from the analysis of the most important metabolic features contributing to the GRN reconstruction, related to either the human or to the mouse organism.

2 Methods

In this section, we first describe how we built the dataset under analysis, from the collection of the gene expression levels for both human and mouse genes to the construction of metabolic features. Then we provide the methodological details of the proposed transfer learning approach. A graphical overview of the proposed approach is shown in Fig. 1.

2.1 Gene expression levels

We collected raw expression data from the Gene Expression Omnibus - GEO (`https://www.ncbi.nlm.nih.gov/geo/`). We considered the platform GPL570 for the human organism and the platform GPL1261 for the mouse organism. We took only control samples to reconstruct the gene regulatory networks, without the potential influence of disease conditions. A complete list of the considered GEO Accession Numbers is reported in Supplementary Table S1.

Quantitatively, for the human organism, we collected $54,675$ probesets, described by 180 samples (that correspond to features in our case): 17 for bone marrow, 37 for brain, 6 for breast, 4 for heart, 7 for liver, 45 for lung, 64 for skin. As for the mouse organism, we collected $45,101$ probesets described by 171 features, distributed as follows according to the organs: 14 for bone marrow, 8 for brain, 10 for breast, 8 for heart, 124 for liver, 4 for lung, and 3 for skin. We processed the raw control samples according to the workflow proposed in the DREAM5 challenge (Marbach *et al.*, 2012). In particular, for each organism, we applied microarray normalization Robust Multichip Averaging (RMA) (Irizarry *et al.*, 2003), considering one batch per organ, through Affymetrix Expression Console Software. Data was background adjusted, quantile normalized, and summarized using median polish. Normalized data was exported as log-transformed expression values. The mapping from Affymetrix probeset IDs to gene IDs was performed through the Affymetrix libraries. Finally, the expression values obtained from multiple probesets mapping to the same gene were aggregated through the arithmetic mean.

2.2 Metabolic features

To construct the metabolic features, we first filtered out genes with no corresponding HGNC ID (Yates *et al.*, 2016). We also removed all the genes for which we did not find any regulatory information according to the RegNetwork database (Liu *et al.*, 2015). Finally, to obtain an expression fold change for constraining the metabolic model, each gene expression value was normalised against its median value across all the samples.

In order to include the regulatory information into the metabolic features explicitly, we used TRFBA (Motamedian *et al.*, 2017b), which integrates a transcriptional regulatory network and the related-organism genome-scale metabolic model (GSMM). We used Recon2.2 (Swainston

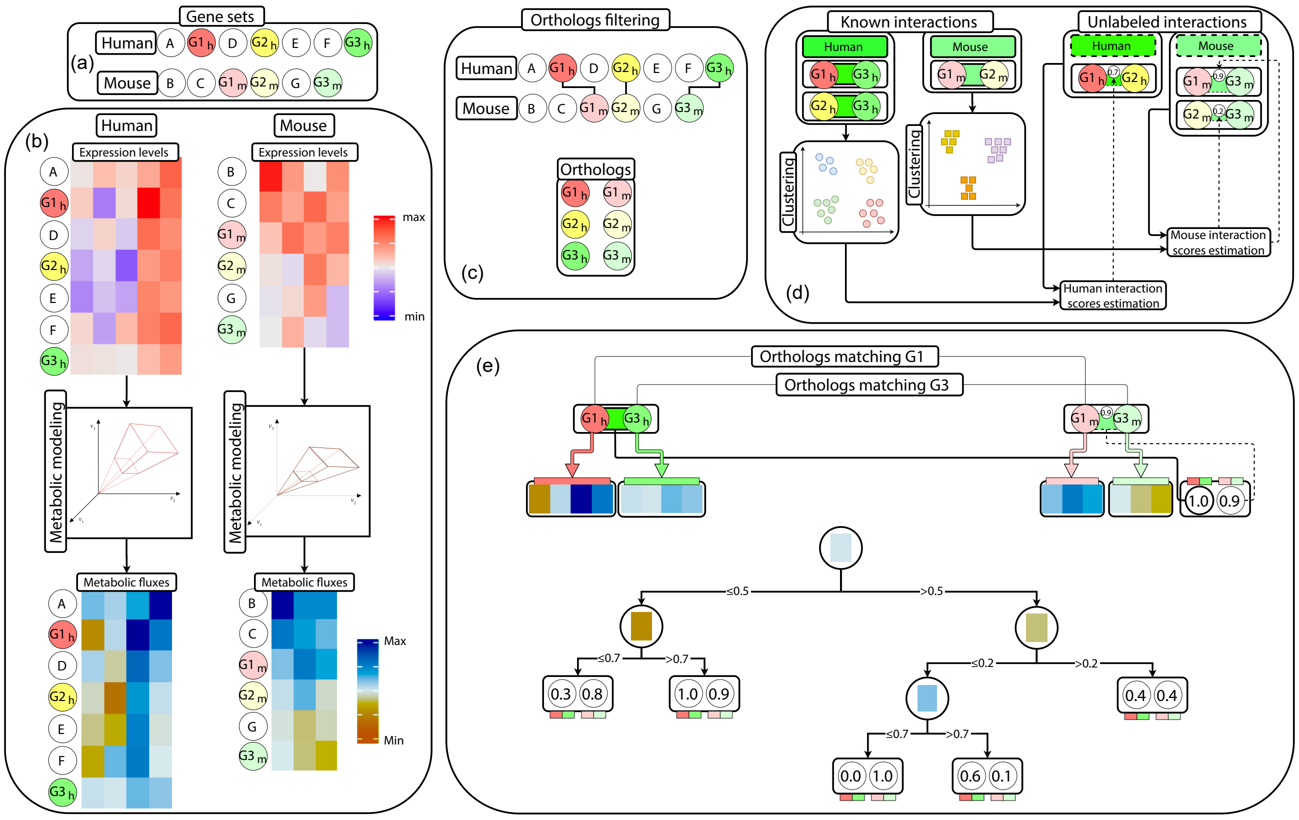

Fig. 1. Starting from the selected gene sets for the human and mouse organisms (a), we compute metabolic fluxes from gene expression levels through genome-scale metabolic modelling of gene knockouts (b) using TRFBA. Genes are then filtered to consider only the subset of orthologous genes for human and mouse (c). For both organisms, we estimate the confidence of existence on unlabelled (i.e. untested) interactions through a clustering-based procedure, and in this way obtain a set of interaction confidence scores (d). Finally, we build multi-target training instances and train a multi-target regression tree (e) to maximize the homogeneity both in the input and in the output spaces, between gene regulations of both human and mouse. The values in the circles of the regression tree represent the prediction (for the human and for the mouse organisms) provided to a gene pair falling into a specific leaf of the regression tree.

et al., 2016) and iMM1415 (Sigurdsson *et al.*, 2010) as the human and mouse metabolic models respectively. The solution selected by TRFBA lies in the feasible solution space defined by the following constraints:

$$\begin{aligned}
&\mathbf{S}\,\mathbf{v} = 0\\
&\mathbf{v}_{lb} \le \mathbf{v} \le \mathbf{v}_{ub}\\
&\sum_{i \in R_j} v_i \le E_j \times C\\
&s_I \times \sum_{r \in G_T} E_r - E_T + U \times w_{I,1} + U \times w_{I,2} \ge -IN_I \qquad (1)\\
&\sum_{r \in G_T} E_r + U \times w_{I,1} - U \le \lambda_I\\
&\sum_{r \in G_T} E_r - U \times w_{I,2} + U \ge \lambda_{I+1},
\end{aligned}$$

where $\mathbf{S}$ is the stoichiometric matrix associated with the species' organism's metabolism, $\mathbf{v}$ is the vector of metabolic flux rates, $\mathbf{v}_{lb}$ and $\mathbf{v}_{ub}$ are the metabolic fluxes lower and upper bounds respectively, R_j is the set of indices corresponding to the reactions which are associated with metabolic gene j, G_T is the set of indices of the regulatory genes of target gene T, E_i indicates the gene expression of gene i, U is a very large number (in our experiments, it was set to the maximum observed expression level multiplied by 5) and s_I, IN_I, $w_{I,1}$, $w_{I,2}$, λ_I and λ_{I+1} are parameters computed directly by the method from the gene expression levels (Motamedian *et al.*, 2017b). The hyperparameter C, used to convert the expression levels of the genes to the upper bounds of the reactions, was set to 0.00014 as suggested by Motamedian *et al.* (2017a). All the other parameters of TRFBA and the boundary constraints for the metabolic models were left to the default values. Therefore, TRFBA adds to the stoichiometric matrix of a GSMM further reactions representing the transcriptomic regulations among the genes, which we exploited by computing the single gene-knockouts and the resulting metabolic fluxes via FBA (Palsson, 2015) for all the genes in the datasets. This was performed for both organisms, obtaining a flux distribution for each knockout.

To account for the tolerance of the solver Gurobi, we then eliminated all the obtained fluxes whose value was lower than 10^{-7} for all the samples, and applied a principal component analysis (PCA) to reduce the dimensionality of the flux distributions. In both cases we retained an explained variance >99%, obtaining 250- and 150-dimensional features for the human and mouse samples, amounting to 1.7% and 0.92% of the original features respectively. These steps were conducted using the COBRA toolbox (Heirendt *et al.*, 2019) in Matlab R2017b.

2.3 Transfer learning for the reconstruction of the human GRN from metabolic features

We here describe our transfer learning approach for the reconstruction of the human GRN, which also exploits the information conveyed by the mouse organism. Our approach learns a model that is able to predict a score in $[0, 1]$, representing the degree of certainty about the existence of a given regulation between two genes. The synergies among the two considered organisms are captured by resembling to a multi-target prediction model, which aims at predicting the existence of a given regulation between two genes in the two organisms simultaneously. Although predicting the existence of a given regulation for the mouse organism is not of specific interest in this study, this strategy allows us to exploit possible correlations between the organisms not only in the input space, but also in the output space (Levatic *et al.*, 2017).

Methodologically, we focused on *orthologous* genes, i.e. different genes of the human and mouse organisms that originated from a single common ancestor gene. Each possible pair of orthologous genes corresponds to a unit of analysis for the predictive task at hand, namely to a possible regulation activity between the two genes. The descriptive attributes of a gene pair correspond to the concatenation of principal component features calculated from flux rates obtained after the respective single-gene knockouts. On the other hand, the value of each target attribute (i.e., the degree of certainty of the existence of such regulation, in the human and in the mouse organisms, respectively) was set to 1.0 if the corresponding gene regulation was experimentally validated according to the BioGRID database (Stark *et al.*, 2006), or estimated through a clustering-based solution (Mignone *et al.*, 2020a) if such regulation has not yet been studied (i.e., it is an unlabelled example). This setting corresponds to the so-called Positive-Unlabelled setting, that is a subclass of the semi-supervised setting as well as a different way to model a one-class classification task (Kaufmann *et al.*, 2020). We note that, for the descriptive attributes of each gene pair, one can in principle compute a flux distribution after a double gene-knockout for the pair, rather than concatenating single-gene knockout fluxes; however, this would require prohibitive computational resources for the dataset and metabolic model at hand (several years of computational time), but it could be a viable approach for smaller models.

Specifically, the known regulations were grouped into clusters, whose number was optimized via a silhouette analysis (Rousseeuw, 1987). The value of the target attributes for unlabelled pairs of genes was then estimated according to the similarity with their closest cluster, computed on the descriptive attributes. Formally, given the descriptive feature vectors $x_h \in \mathbb{R}^p$ (for the human organism) and $x_m \in \mathbb{R}^q$ (for the mouse organism) for the same gene pair, we computed the value of the target variables t_h (for the human organism) and t_m (for the mouse organism) to use during the training of the predictive model, as follows:

$$\begin{aligned} t_h(x_h) &= \max_{c \in C_h} sim_p(x_h, cent(c)) \\ t_m(x_m) &= \max_{c \in C_m} sim_q(x_m, cent(c)), \end{aligned} \tag{2}$$

where C_h and C_m are the sets of clusters identified for the human and mouse organisms, respectively; $cent(c)$ is the feature vector of the centroid of the cluster c; sim_k: $\mathbb{R}^k \times \mathbb{R}^k \rightarrow [0,1]$ is a vector similarity function working on arbitrary k-dimensional vectors, based on the Euclidean distance after applying a min-max normalization (in the range $[0,1]$) to all the descriptive features. Formally, $sim_k(a,b) = 1 - 1/k \cdot \sqrt{\sum_{i=1}^{k}(a_i - b_i)^2}$. In this way, we exploited both the information on verified regulations and the information conveyed by a large set of unlabelled examples, according to their similarity with respect to labelled examples.

Finally, we built a predictive model in the form of a multi-target regression tree, by exploiting the system CLUS (Levatic *et al.*, 2018), which is based on the predictive clustering framework. Predictive clustering approaches appear adequate to solve the task at hand, since they have proven to be generally effective in detecting different kinds of autocorrelation phenomena (Corizzo *et al.*, 2019), including network autocorrelation phenomena usually exhibited by data organized in network structures (Pio *et al.*, 2018; Serafino *et al.*, 2018).

The multi-target regression tree was built via a standard procedure for the top-down induction of regression trees, where the tests of the internal nodes are greedily chosen by considering the reduction of variance achieved by partitioning the examples according to this test. In our case, the model aims to reduce the variance of both target attributes t_h and t_m. More formally, for a given internal node of the tree under construction, it aims to maximize the reduction of the average variance over the target attributes due to the split, namely

$$Var_X(t_h, t_m) - (Var_{X'}(t_h, t_m) + Var_{X''}(t_h, t_m)), \tag{3}$$

where X, X', X'' are the sets of examples in the parent, left child and right child nodes, respectively, and $Var_Z(t_h, t_m) = \frac{Var_Z(t_h) + Var_Z(t_m)}{2}$ is the average variance on the target attributes t_h and t_m, computed over the set of examples Z. As a result, we maximized the homogeneity of the defined subsets of examples, that also depends on the correlations, both in the input and in the output spaces, between gene regulations of both the human and the mouse organisms.

3 Results and Discussion

The final network under consideration consists of 512,576 possible interactions. Among them, 507,656 are unlabelled, while 4,920 are labelled/known interactions from BioGRID. Therefore, the proportion of labelled:unlabelled interactions is ∼1:100.

We compare the results obtained by our framework based on metabolic features, hereafter referred to as TRANSFER, with those achieved by two different settings:

- **Expression Levels**. We adopt the same workflow proposed in this paper, but directly using the expression level features instead of metabolic features. This setting allows us to evaluate the actual contribution provided by metabolic features.
- **NOTRANSFER**. We only exploit features related to the genes of the human organism. This setting allows us to evaluate the contribution provided by information conveyed by the mouse organism as well as the effectiveness of the proposed transfer learning solution.

The experiments were performed through 10-fold cross-validation. In particular, each fold consists of 9/10 positive examples for training and 1/10 positive examples for testing, while all the unlabelled examples are used for both training and testing purposes. Therefore, coherently with the *semi-supervised transductive* setting (Ji *et al.*, 2010; Ma *et al.*, 2020), at training time the methods know the examples for which they have to make a prediction, i.e., they may already observe and exploit the value of descriptive attributes, but not the actual value of the target attributes. We note that the confidence scores estimated by our method are not adopted to define a ground truth for unlabelled examples, but only as an intermediate step for the construction of the multi-target regression tree.

The results were evaluated in terms of recall@k (r@k), the area under the recall@k curve (AUR@K), the area under the ROC curve (AUROC) and the area under the precision-recall curve (AUPR). We note that, while r@k and AUR@K do not introduce any bias on the existence of a regulation activity on unlabelled gene pairs, the computation of the AUROC and AUPR requires considering the unlabelled examples as negative examples.

In Fig. 2 we show the measured recall@k in the range $[0, 1\%]$, that is the range of the top-1% most reliable interactions returned by all the approaches considered. Our results show that the adoption of metabolic fluxes is beneficial, with respect to directly adopting the raw gene expression levels, both when exploiting the knowledge coming from the mouse organism (TRANSFER) and when ignoring such additional information (NOTRANSFER). Specifically, such an improvement amounts to 6.6% in the case of NOTRANSFER and to 8.73%

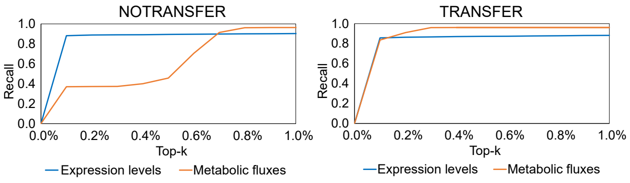

Fig. 2. Recall@k measured in the range $[0, 1\%]$ for the reconstruction of the human GRN, by considering different sets of features. The NOTRANSFER approach does not exploit data of the mouse organism, while the TRANSFER approach exploits also the mouse GRN knowledge.

in the case of TRANSFER, when observing the recall@1%. Moreover, it is noteworthy that, in the TRANSFER setting, we identify existing gene regulations much earlier in the returned ranked list of interactions. Specifically, we identify 96% of the known interactions of the testing set in the top-0.3% interactions returned in the case of the TRANSFER setting, whereas we need to consider 0.8% of the list of the returned interactions to identify the same amount of known interactions in the NOTRANSFER setting. This behaviour emphasizes that the knowledge coming from the mouse organism can fruitfully be exploited to improve the accuracy of the reconstruction of the human GRN.

A more comprehensive overview is reported in Fig. 3, where we show boxplots representing AUR@K, AUROC and AUPR measured over the 10 folds of the cross-validation. These show that the predictive models trained via metabolic fluxes can better exploit the mouse gene regulation knowledge leading to more stable predictive models (i.e., with a lower variance observed over different folds of the 10-fold CV). Furthermore, the area under all the considered curves is higher and more stable when adopting the metabolic fluxes in combination with the TRANSFER setting. Conversely, when adopting metabolic fluxes in the NOTRANSFER setting, we observe worse results with respect to directly using expression levels. This phenomenon indicates that the metabolic fluxes of the human organism alone are not able to describe the regulatory activities as well as expression levels, but the exploitation of mouse and human metabolic fluxes in combination provides our framework with a significant advantage, leading to the best overall results. This observation confirms that the proposed workflow, which synergically exploits metabolic fluxes and the knowledge of the mouse GRN, provides significant advantages in terms of the quality of the reconstruction of the human GRN.

To understand the contribution provided by human and mouse metabolic features in human GRN reconstruction, we performed additional experiments in the TRANSFER and NOTRANSFER settings. Specifically, we considered the approach proposed by Petković *et al.* (2017), based on the evaluation of the (negative) effect of noise. We purposely introduced noise in a given feature by randomly permuting its values over the examples, evaluating the effect on the predictive performance of the tree: the greater the performance degradation, measured through the relative increase of the predictive error, the higher is the importance/contribution of the feature. We produced a descending ranking of the features for each fold of the 10-fold cross-validation, and analyzed the average ranks (see Supplementary Table S2 for the computed average ranks).

As shown in Fig. 4(c), the metabolic features related to the mouse organism dominate the upper half of the ranking and are therefore assigned a higher relevance than those related to the human, in the setting TRANSFER. Conversely, when directly using gene expression levels, many features from human retain a high relevance when combined with those from mouse (see Supplementary Table S2). This finding further confirms the advantage provided by the adoption of our transfer learning technique on GSMM-derived information.

Further, a consistent number of metabolic features (295/800) present a relevance score equal to zero, as opposed to the more gradual decline in gene expression feature relevance. This is in line with previous experimental results from another data integration task, where metabolic features displayed a highly skewed relevance distribution compared to transcriptomic ones (Culley *et al.*, 2020). A possible explanation is given by the structure of metabolic networks and by the method used to estimate its activity, which is based on a linearly constrained MILP problem that generates collinearity and redundancy among the features.

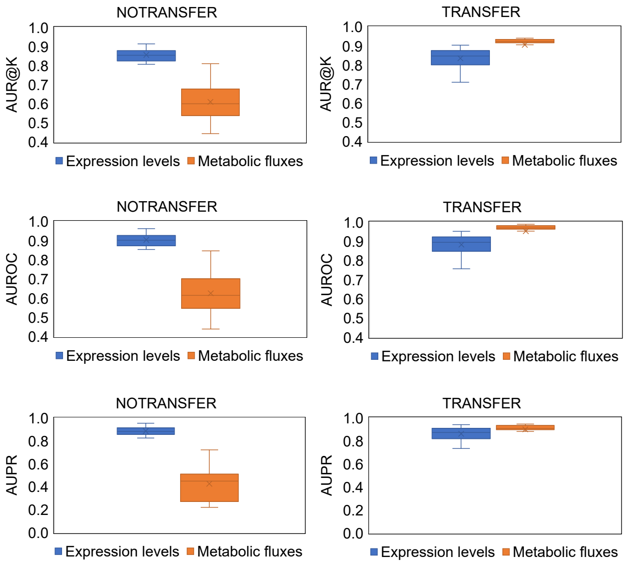

Fig. 3. Boxplots for the 10 folds of the human GRN reconstruction task. Each row corresponds to a measure, i.e., AUR@K, AUROC, AUPR, respectively, measured in the range [0, 1%] of the top-k ranked interactions. Each column corresponds to a learning setting, i.e., without and with the exploitation of the mouse GRN knowledge, respectively.

Consistently, the addition of mouse-related features impacts the importance of human-related features to a varying degree depending on their type. When comparing the TRANSFER and NOTRANSFER settings, human metabolic features have indeed an average importance difference of 2.12±2.80, whereas for human transcriptomic features such difference is 2.99±1.25. In other words, human gene expression features that are considered poorly (or highly) relevant in the NOTRANSFER scenario have on average a higher chance to be considered more (or less) relevant in the TRANSFER setting – and by a larger extent – as compared to human metabolic features. However, the difference in importance for the latter is highly variable and reaches the highest values. The addition of mouse-related features therefore appears to drastically change the learned model when using metabolic features.

To further characterise our results, we inspected the metabolic pathways associated with the most relevant reactions adopted in the construction of the metabolic features. We conducted a flux enrichment analysis using the MATLAB Bioinformatics Toolbox on the subset of reactions which, for each organism in the two experimental settings (TRANSFER and NOTRANSFER), had been given a weight above the 90th percentile. The weight for the j-th reaction was computed as

$$\theta_j = \sum_i |l_{ij} \times \sigma_i^2 \times (rank_{1j} + rank_{2j})|, \quad (4)$$

where l_{ij} is the linear coefficient of the j-th feature/reaction with respect to the i-th principal component deriving from the PCA (adopted to generate the metabolic features), σ_i^2 is the variance explained by the i-th principal component, while $rank_{1j}$ and $rank_{2j}$ are the rankings of the j-th feature, computed using the approach proposed by Petković *et al.* (2017), when considered in the first and second position, respectively, in the gene pair.

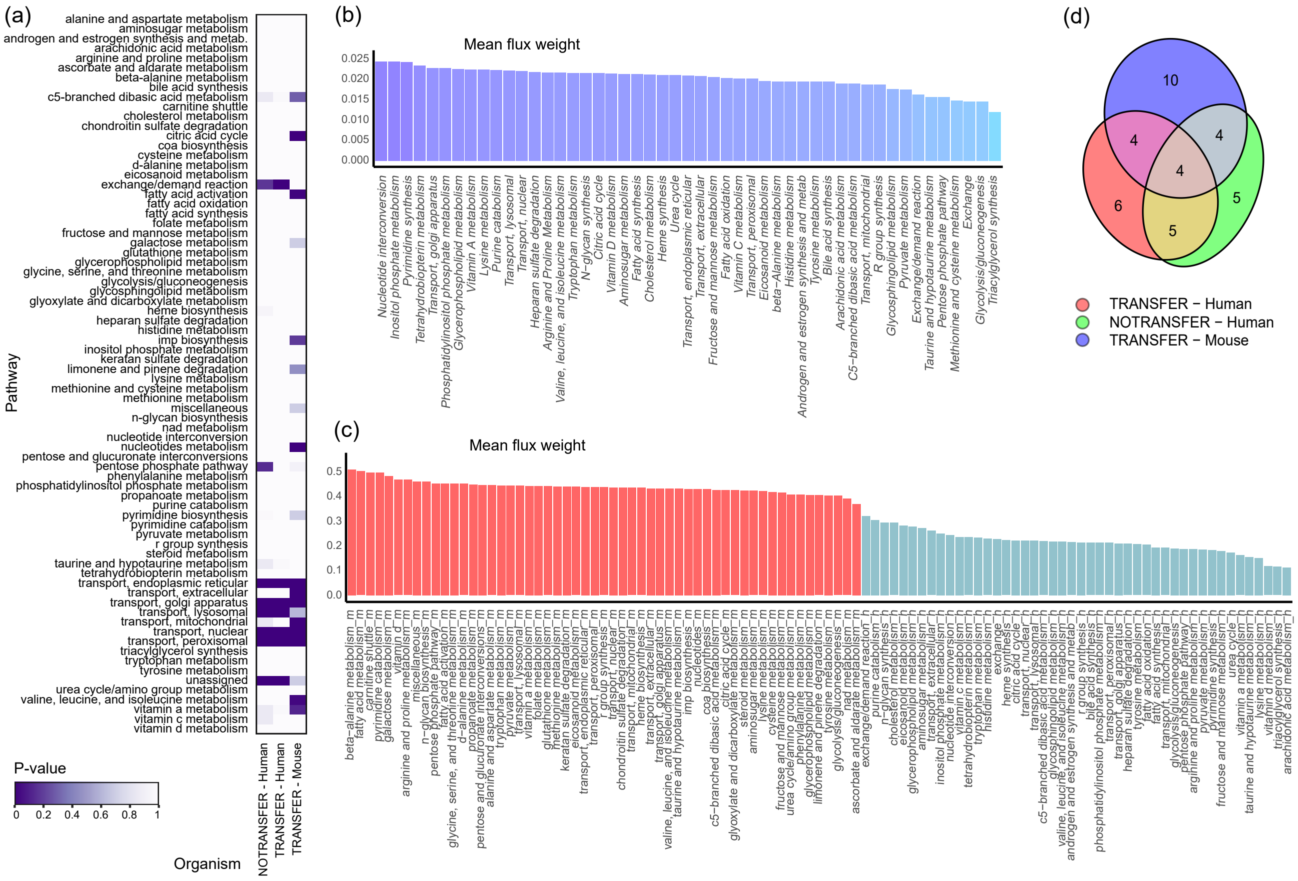

Fig. 4. (a) Enrichment p-values (corrected through the Benjamini-Hochberg procedure for multiple hypothesis testing) for the pathways assigned to the 10% most relevant metabolic features in the three experimental settings considered. (b) Mean flux weight across pathways for the human metabolic features used in the setting NOTRANSFER. (c) Mean flux weight across pathways for the human (blue) and mouse (red) metabolic features used in the setting TRANSFER. (d) Euler-Venn diagram that summarises the overlap in terms of biological pathway enrichment (pathways with associated corrected p-value ≤ 0.05) for the 10% most relevant metabolic features in the three considered settings.

From these values, we computed the average flux weight for each metabolic pathway as the average weight of its reactions.

As shown in Fig. 4(a)(d), the number of enriched pathways (associated p-value ≤ 0.05, corrected through the Benjamini-Hochberg procedure for multiple hypothesis testing) is higher for the metabolic features of the mouse, while it is almost equal for the human ones. Indeed, reactions enriched in the human features employed when building the model without the mouse features (NOTRANSFER-Human) were all included in the pool of enriched reactions from the human features used in the TRANSFER setting. In particular, in this setting, the enrichment also includes exchange/demand reactions (p-value > 0.05 for the NOTRANSFER-Human features), indicating that adding features from a different organism increased the importance of the features associated with internal production/consumption reactions and extracellular/intracellular transport reactions. Conversely, mouse features share all the transport pathways of the human ones, except for that relating to lysosomal transport, and also encompass the pathways associated with the citric acid cycle, nucleotide metabolism, fatty acid activation and the metabolism of leucine, isoleucine and valine (see Fig. 4 (a)).

Overall, these results demonstrated the effectiveness of the proposed approach, which exploits metabolic information coming from two organisms through our transfer learning method. Moreover, the analysis of the contribution of the metabolic features emphasized the new information introduced by the mouse features. We believe that our results pave the way towards the exploitation of knowledge of multiple model organisms - across several omic layers - while reconstructing the GRN of a target organism.

4 Conclusion

We presented a novel method for the reconstruction of the human gene regulatory network that fruitfully exploits the information conveyed by *in silico*-generated metabolic fluxes of both mouse and human organisms. Specifically, we exploit a transfer learning method to capture analogies between the metabolic responses in mouse and human upon simulated deletion of their orthologous genes.

Our results show that metabolic features, computed from gene expression levels and metabolic modelling, improve the performance and the stability of the trained predictive models when exploited in combination with our transfer learning approach. This emphasizes that the underlying regulatory patterns are better captured when (both known and possible) gene regulations are described through metabolic features, computed through genome-scale metabolic model simulations, on both the human and the mouse organisms. To our knowledge, this work is the first attempt to exploit metabolic features and a transfer learning approach for the reconstruction of the human GRN, and our results support the adoption of the developed method as a state-of-the-art tool for solving this task.

As future work, we aim to design a multi-source approach to capture possible dependencies among multiple organisms and to simultaneously reconstruct their GRNs, even when the knowledge about their orthologous genes is limited. In conjunction with multi-omic integration strategies, this could lead to refined GRN reconstructions, thus expanding the current knowledge on the biological mechanisms of metabolic regulation.

Funding

GP acknowledges the support of the Ministry of Universities and Research through the project "Big Data Analytics", AIM 1852414-1(line 1). GM, GZ and CA acknowledge the support of UKRI Research England's THYME project, and a Children's Liver Disease Foundation Research Grant. PM acknowledges the support of the Apulia Region through the "Research for Innovation - REFIN" initiative (Grant n. 7EDD092A).

References

Ben Guebila, M. and Thiele, I. (2019). Predicting gastrointestinal drug effects using contextualized metabolic models. *PLoS computational biology*, **15**(6), e1007100.

Chandrasekaran, S. and Price, N. D. (2010). Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in escherichia coli and mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, **107**(41), 17845–17850.

Corizzo, R., Pio, G., Ceci, M., and Malerba, D. (2019). DENCAST: distributed density-based clustering for multi-target regression. *J. Big Data*, **6**, 43.

Culley, C., Vijayakumar, S., Zampieri, G., and Angione, C. (2020). A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. *Proceedings of the National Academy of Sciences*, **117**(31), 18869–18879.

Davidson, E. H. (2010). Emerging properties of animal gene regulatory networks. *Nature*, **468**(7326), 911–920.

Gardner, T. S., Di Bernardo, D., Lorenz, D., and Collins, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**(5629), 102–105.

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., *et al.* (2019). Creation and analysis of biochemical constraint-based models using the cobra toolbox v. 3.0. *Nature protocols*, **14**(3), 639–702.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2), 249–264.

Ji, M., Sun, Y., Danilevsky, M., Han, J., and Gao, J. (2010). Graph regularized transductive classification on heterogeneous information networks. In *Machine Learning and Knowledge Discovery in Databases*, pages 570–586. Springer Berlin.

Karlebach, G. and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, **9**(10), 770–780.

Kaufmann, J., Asalone, K., Corizzo, R., Saldanha, C., Bracht, J., and Japkowicz, N. (2020). One-class ensembles for rare genomic sequences identification. In *International Conference on Discovery Science*, pages 340–354. Springer.

Kavvas, E. S., Yang, L., Monk, J. M., Heckmann, D., and Palsson, B. O. (2020). A biochemically-interpretable machine learning classifier for microbial gwas. *Nature communications*, **11**(1), 1–11.

Levatic, J., Ceci, M., Kocev, D., and Dzeroski, S. (2017). Semi-supervised classification trees. *J. Intell. Inf. Syst.*, **49**(3), 461–486.

Levatic, J., Kocev, D., Ceci, M., and Dzeroski, S. (2018). Semi-supervised trees for multi-target regression. *Inf. Sci.*, **450**, 109–127.

Liu, Z.-P., Wu, C., Miao, H., and Wu, H. (2015). Regnetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, **2015**.

Luo, Q., Liu, X., and Yi, D. (2009). Reconstructing gene networks from microarray time-series data via granger causality. In J. Zhou, editor, *Complex Sciences*, pages 196–209, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ma, Y., Bai, S., An, S., Liu, W., Liu, A., Zhen, X., and Liu, X. (2020). Transductive relation-propagation network for few-shot learning. In *IJCAI 20*, pages 804–810.

Magazzù, G., Zampieri, G., and Angione, C. (2021). Multimodal regularised linear models with flux balance analysis for mechanistic integration of omics data. *Bioinformatics*.

Marbach, D., Costello, J. C., Küffner, R., *et al.* (2012). Wisdom of crowds for robust gene network inference. *Nat Methods*, **9**(8), 796–804.

Mignone, P., Pio, G., D'Elia, D., and Ceci, M. (2020a). Exploiting transfer learning for the reconstruction of the human gene regulatory network. *Bioinform.*, **36**(5), 1553–1561.

Mignone, P., Pio, G., Džeroski, S., and Ceci, M. (2020b). Multi-task learning for the simultaneous reconstruction of the human and mouse gene regulatory networks. *Scientific Reports*, **10**, 22295.

Motamedian, E., Taheri, E., and Bagheri, F. (2017a). Proliferation inhibition of cisplatin-resistant ovarian cancer cells using drugs screened by integrating a metabolic model and transcriptomic data. *Cell proliferation*, **50**(6), e12370.

Motamedian, E., Mohammadi, M., Shojaosadati, S. A., *et al.* (2017b). TRFBA: an algorithm to integrate genome-scale metabolic and transcriptional regulatory networks with incorporation of expression data. *Bioinform.*, **33**(7), 1057–1063.

Nielsen, J. (2017). Systems biology of metabolism: a driver for developing personalized and precision medicine. *Cell metabolism*, **25**(3), 572–579.

Occhipinti, A., Hamadi, Y., Kugler, H., Wintersteiger, C., Yordanov, B., and Angione, C. (2020). Discovering essential multiple gene effects through large scale optimization: an application to human cancer metabolism. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Palsson, B. Ø. (2015). *Systems Biology: Constraint-Based Reconstruction and Analysis*. Cambridge University Press.

Petković, M., Džeroski, S., and Kocev, D. (2017). Feature ranking for multi-target regression with tree ensemble methods. In *Proc. of DS 2017*, pages 171–185.

Pio, G., Serafino, F., Malerba, D., and Ceci, M. (2018). Multi-type clustering and classification from heterogeneous networks. *Inf. Sci.*, **425**, 107–126.

Pio, G., Ceci, M., Prisciandaro, F., and Malerba, D. (2020). Exploiting causality in gene network reconstruction based on graph embedding. *Mach. Learn.*, **109**(6), 1231–1279.

Richelle, A., Kellman, B. P., Wenzel, A. T., Chiang, A. W., Reagan, T., Gutierrez, J. M., *et al.* (2020). What does your cell really do? Model-based assessment of mammalian cells metabolic functionalities using omics data. *bioRxiv*.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Comput. Appl. Math.*, **20**, 53 – 65.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, **4**(1).

Schlitt, T. and Brazma, A. (2007). Current approaches to gene regulatory network modelling. *BMC bioinformatics*, **8**(S6), S9.

Serafino, F., Pio, G., and Ceci, M. (2018). Ensemble learning for multi-type classification in heterogeneous networks. *IEEE Trans. Knowl. Data Eng.*, **30**(12), 2326–2339.

Sigurdsson, M. I., Jamshidi, N., Steingrimsson, E., Thiele, I., and Palsson, B. Ø. (2010). A detailed genome-wide reconstruction of mouse metabolism based on human recon 1. *BMC systems biology*, **4**(1), 140.

Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, **34**(Database-Issue), 535–539.

Swainston, N., Smallbone, K., Hefzi, H., Dobson, P. D., Brewer, J., Hanscho, M., Zielinski, D. C., Ang, K. S., Gardiner, N. J., *et al.* (2016). Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*, **12**(7), 109.

Tonkovic, P., Kalajdziski, S., Zdravevski, E., Lameski, P., Corizzo, R., *et al.* (2020). Literature on metagenomic classification: scoping review on machine learning trends in metagenomics. *Biology*.

Töpfer, N., Kleessen, S., and Nikoloski, Z. (2015). Integration of metabolomics data into metabolic networks. *Frontiers in plant science*, **6**, 49.

Vijayakumar, S., Conway, M., Lió, P., and Angione, C. (2018). Seeing the wood for the trees: a forest of methods for optimization and omic-network integration in metabolic modelling. *Briefings in bioinformatics*, **19**(6), 1218–1235.

Wang, Z., Danziger, S. A., Heavner, B. D., Ma, S., Smith, J. J., Li, S., *et al.* (2017). Combining inferred regulatory and reconstructed metabolic networks enhances phenotype prediction in yeast. *PLoS computational biology*, **13**(5), e1005489.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., *et al.* (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, **42**(D1), D1001–D1006.

Yang, J. H., Wright, S. N., Hamblin, M., McCloskey, D., Alcantar, M. A., *et al.* (2019). A white-box machine learning approach for revealing antibiotic mechanisms of action. *Cell*, **177**(6), 1649–1661.

Yates, B., Braschi, B., Gray, K. A., Seal, R. L., Tweedie, S., and Bruford, E. A. (2016). Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic acids research*, page gkw1033.

Ye, Y., Li, S.-L., and Wang, S.-Y. (2018). Construction and analysis of mrna, mirna, lncrna, and tf regulatory networks reveal the key genes associated with prostate cancer. *PloS one*, **13**(8), e0198055.

Yeang, C.-H. and Vingron, M. (2006). A joint model of regulatory and metabolic networks. *BMC bioinformatics*, **7**(1), 332.

Yurkovich, J. T. and Palsson, B. O. (2015). Solving puzzles with missing pieces: the power of systems biology. *Proceedings of the IEEE*, **104**(1), 2–7.

Zampieri, G., Vijayakumar, S., Yaneske, E., and Angione, C. (2019). Machine and deep learning meet genome-scale metabolic modeling. *PLoS computational biology*, **15**(7), e1007084.

Zou, M. and Conzen, S. D. (2005). A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, **21**(1), 71–79.