Cell-cell communication

University of Catania

Finding edge weights and building the iteration algorithm

Locicero Giorgio
giorgio.locicero@phd.unict.it

January 11, 2023

1 Introduction

The whole process is iterative. The pre-embedding, post-embedding, and pre-CCI steps are repeated until convergence is reached either by:

- Termination condition
- Scores are ϵ limited during the iteration and do not variate too much
- Error (to be treated accordingly)

1.1 pre-embedding data

- expression profile for the single cells
- annotation data for the single cells (typed cell, origin, metadata for the origin), derivable from the expression profiles or other forms of derivation (manual labeling with images, image processing, origin information)
- logfold-change(derived from expression profiles and annotation data for groups) over:
  - **intra-sample** over the same sample between different cell types
  - **inter-sample** over different samples between different patients
  - **complete** over different samples aggregated (with pseudo-bulk or similar aggregation methods), between different cell types, or between different patients

1.2 post-embedding data

- **Perturbation scores** for all the genes in the metapathway, different for every cell type. Not accounting yet for the cell communication
- **Pathway overall perturbation score** for all the pathways, different for every cell type. Again, not accounting yet for the cell communication

1.3 pre-CCI data

- All the previously seen data from pre-embedding and post-embedding steps
- LR list from database
- Other sources of data that could be used for the estimation of the CCI interactions

During this step, the data will be treated to get the interaction weights that will be used for the iterative algorithm to get the perturbation intra and inter pathway, between different cells. The edge weights could be:

- The **probability** of interaction, either conditional or fuzzy
- The **Normalized score** for the interaction (like ICELLNET or cell2cell), this normalized score can be derived from:
  - a **binary** score function, mainly thresholding, differential combination or hypothesis testing
  - a **continous** score function, either derived from gene co-expression or **expression product**
- The **co-expression/correlation** between pairs and cells

There is also the need to understand if cells could be treated as virtual nodes to also get an interactions score for the cell as well(and an interaction weight before the algorithm as well). Also, the use of a methodology similar to that of **STEPMINER** and **BOOLEANET** can be useful since interactions and implications between LR pairs need to be found as well.

1.3.1 Assigning edge weights to ligand-receptors

One of the most simple and already tested cases could be to assign a weight based upon the gene co-expression (WGCNA), but, at the same time, the interactions of ligand receptors should not be treated as gene interactions but as protein interactions.

Also, I do not know if the metapathway has the ligand-receptor edges as well, but i do not think so since those could create cycles and are also not part of common metapathways.

1.4 post-CCI data

- **Perturbation score** for the genes, including ligands and receptors, along augmented pathways as well(miRNA, REACTOME, etc...)
- **Interaction score** between cells types

2 Things to take into account

2.1 Cycles

Cycles are a tricky thing to treat, especially when using recursive or iterative algorithms. If convergence is guaranteed, cycles can be ignored since they can be treated as **strongly**

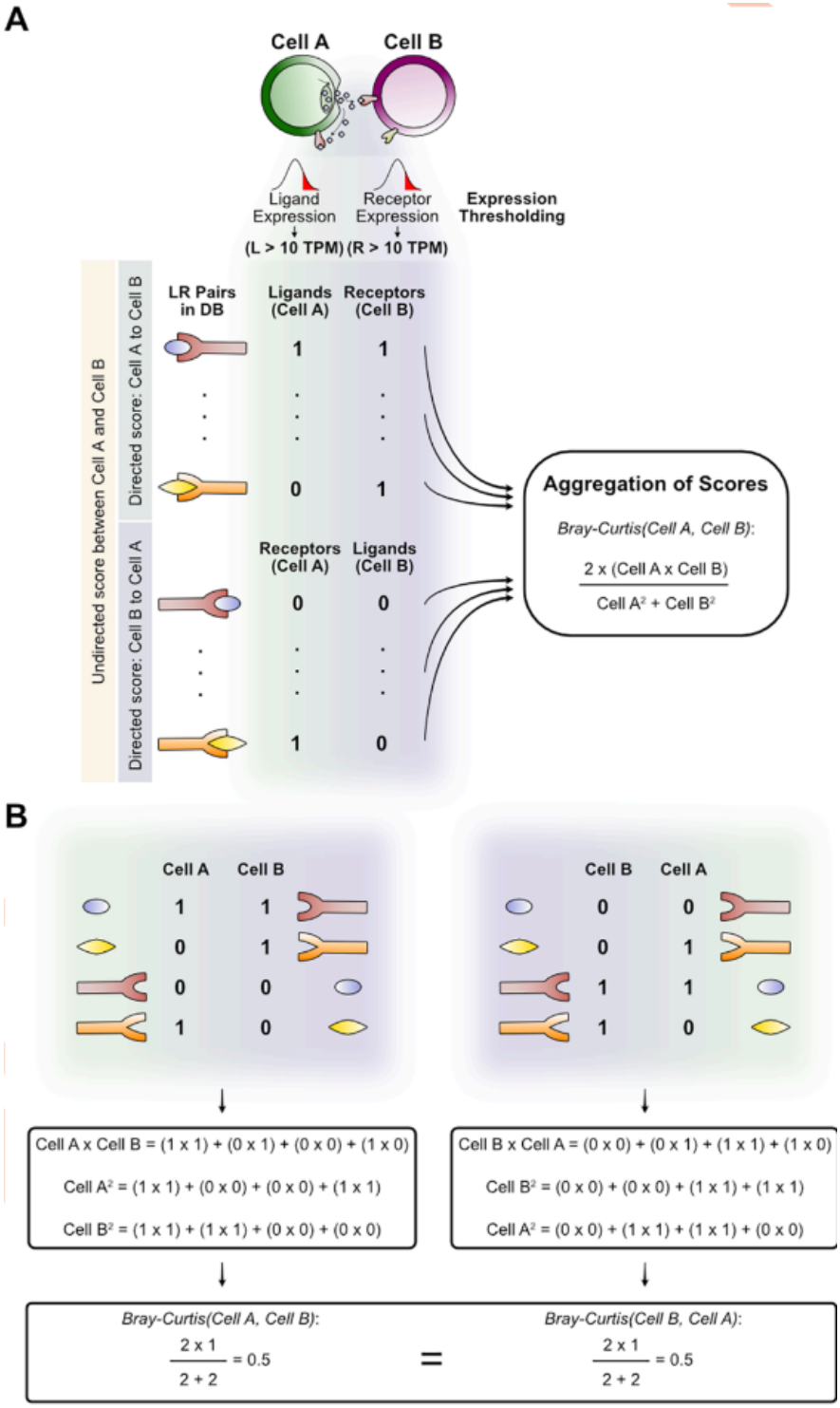

Figure 1: cell2cell interaction score

connected components and can eventually be aggregated into a single component for the network.

2.2 Short vs long-range communication

2.3 Known LR pairs vs unknown but significative pairs

The significative LR pairs are identified during the **pre-CCI** step.

2.4 Testing the interactions found

Done with:

- Comparing known interactions from databases(ICELLNET)
- Comparing interaction data with Spatial transcriptomics if available

3 C2C-SEPIA

Cell-to-cell single-cell enriched pathway impact analysis(C2C-SEPIA) is a framework that uses pathways and enriched pathways to incorporate CCC into the computation of the perturbation in different genes and in the whole pathways from single-cell transcriptome data. For every type inferred or available from the single-cell expression profiles, the method estimates the interaction between cells and the perturbation in the network of pathways(intra-cellular perturbation) and cells(inter-cellular perturbation).

The gist of the new methodology is to have virtual nodes that represent cell types in the metapathway of a defined cell type, the communication between cells is modeled by making these virtual nodes interact with the pathways with the use of Ligand-Receptors between cell types.

A virtual node is present in the metapathway if at least an LR pair is co-expressed, more formally:

Definition 3.1. (Typed virtual input node presence in the metapathways). Given two metapathways, $M(t_d)$ related to the cell type t_d and $M(t_s)$ related to the cell type t_s, A virtual node $v_{t_s \to t_d}$ for the cell-type t_s is present in the metapathway $M(t_d)$ if at least a ligand l_s related to a gene $g_{M(t_s)}(l_s)$ in the metapathway $M(t_s)$ and the receptor r_d related to a gene $g_{M(t_d)}(r_d)$ in the metapathway $M(t_d)$ are positively co-expressed (positively correlated) between each other.

In a similar way, a virtual output node can be defined:

Definition 3.2. (Typed virtual output node presence in the metapathways). Given two metapathways, $M(t_d)$ related to the cell type t_d and $M(t_s)$ related to the cell type t_s, A virtual node $v_{t_d \leftarrow t_s}$ for the cell-type t_s is present in the metapathway $M(t_s)$ if at least a ligand l_s related to a gene $g_{M(t_s)}(l_s)$ in the metapathway $M(t_s)$ and the receptor r_d related to a gene $g_{M(t_d)}(r_d)$ in the metapathway $M(t_d)$ are positively co-expressed (positively correlated) between each other.

Definition 3.3. (Typed virtual output set for a cell type). The virtual output set of a metapathway $M(t)$ related to the cell type t is the set $V_{\text{out}}(t) = \{v_{t_d \leftarrow t} | t_d \in CT \setminus t\}$ Where CT is the set of cell types available for the cells(from inference or from manual labeling) and associated with the cell profiles.

Definition 3.4. (Typed virtual input set for a cell type). The virtual input set of a metapathway $M(t)$ related to the cell type t is the set $V_{\text{in}}(t) = \{v_{t_s \rightarrow t} | t_s \in CT \setminus t\}$

In the case of the same cell-type communication, the framework doesn't consider this case since it will introduce non-convergence for the perturbation in the metapathway but it will be considered in the future when the method is more refined. Also, the ligand-receptor pairs in the metapathway are already considered(???)

With these virtual nodes added to the metapathways (for every cell type), the new graph nodes is called **augmented metapathway**, the set of nodes in the metapathway will be the virtual nodes for the output and input and the already present genes in the metapathway, formally:

Definition 3.5. (Augmented metapathway nodes). Given a metapathway $M(t) = (V(t), E(t))$ related to the cell-type t and two sets of typed virtual nodes for input $V_{\text{in}}(t)$ and for output $V_{\text{out}}(t)$, the set of nodes in the augmented metapathway is defined as

$$V^*(t) = V(t) \cup V_{\text{in}}(t) \cup V_{\text{out}}(t)$$

One thing to consider is that only one typed virtual input node and one typed virtual output node can be added to a metapathway for every cell type(one for the ligands interactions, and one for the receptors' interaction).

These virtual nodes and their edges code the interaction between the cell types, these edges have been chosen to be from the virtual node(input or output) to receptors(in case of the virtual inputs) and ligands (in case of the virtual outputs), formally:

Definition 3.6. (Typed virtual input interaction). The interaction between a typed virtual input node $v_{t_s \rightarrow t} \in V_{in}(t)$ and the genes in the metapathway $M(t)$ related to the cell type t is present in the augmented metapathway if and only if the receptor gene $g_r \in Receptors(t)$ in the single-cell gene profiles for t is co-expressed with the ligand gene in the single-cell gene profiles for the type t_s. That is

$$(v_{t_s \rightarrow t}, g_r, 1) \in E_{\text{in}}(t) \iff g_r \in Receptors(t), \exists g_l \in V(t_s) | g_l \in Ligands(t_s), \text{co-expressed}(g_l, g_r)$$

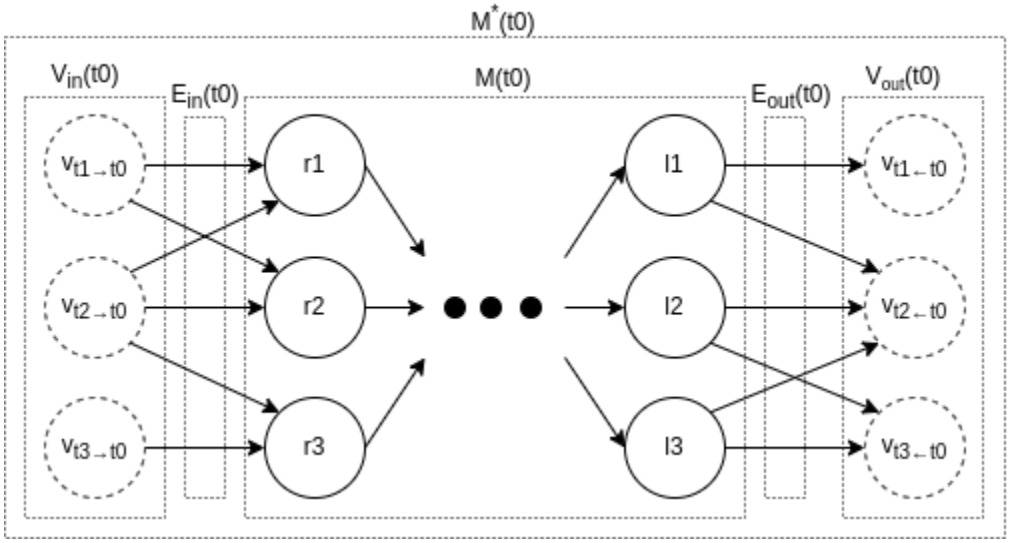

Figure 2: visualization of the augmented metapathway graph, an example

In the same way, the typed virtual output interaction can be defined.

Definition 3.7. (Typed virtual output interaction). The interaction between genes in the metapathway $M(t)$ related to the cell type t and a typed virtual output node $v_{t_d \leftarrow t} \in V_{out}(t)$ is present in the augmented metapathway if and only if the ligand gene in the single-cell gene profiles for t is co-expressed with the receptor gene in the single-cell gene profiles for the type t_d. That is

$$(g_l, v_{t_d \leftarrow t}, 1) \in E_{\text{out}}(t) \iff g_l \in Ligands(t), \exists g_r \in V(t_d) | g_r \in Receptors(t_d), \text{co-expressed}(g_r, g_l)$$

All the interactions will have a weight of 1 to simplify the model.

The corresponding augmented metapathway will be $M^*(t) = (V^*(t), E^*(t)), V^*(t) \supseteq V(t), E^*(t) = E(t) \cup E_{\text{in}}(t) \cup E_{\text{out}}(t)$.

A visualization of the augmented metapathway graph and its component can be seen in figure 2

There is a clear definition for virtual nodes in metapathways because every metapathway associated with a cell type needs to have a different set of virtual nodes, but a typed virtual input node $v_{t_s \rightarrow t} \in V(t)$ in the augmented metapathway $M^*(t)$ is associated with $v_{t \leftarrow t_s} \in V(t_s)$ in the augmented metapathway $M^*(t_s)$ with the following formula following the iterative algorithm:

$$v_{t_s \rightarrow t}^{(n)} = \begin{cases} 0 & n = 0 \\ v_{t \leftarrow t_s}^{(n-1)} & n > 0 \end{cases} \tag{1}$$

After every iteration of the algorithm, the values for the virtual input nodes are updated with the previous definition.

The metapathway $M(t)$ is associated with a matrix $W \in \mathbb{R}^{|V(t)|x|V(t)|}$ that is used in the original framework in the following way:

Definition 3.8. (Perturbation values of the original framework). Given an input vector $Input \in \mathbb{R}^{|V(t)|}$ that represents the log2fold-change passed as an input for every node in the metapathway, a matrix $W \in \mathbb{R}^{|V(t)|x|V(t)|}$ that represents the edges weight in the metapathway, an identity matrix I, the perturbation vector $P \in \mathbb{R}^{|V(t)|}$ is computed with the following equation:

$$(I - W^T) * P = Input \tag{2}$$

The modified equation that will be used to compute the perturbation is formed by the same matrix used for the original algorithm plus the additional edges from the augmented metapathway, that is $W^* \in \mathbb{R}^{|V^*(t)|x|V^*(t)|}$, the modified equation that will be used in the iteration algorithm is defined as follows:

$$(I - W^{*^T}) * P^{*(n)} = \begin{cases} Input^* & n = 0 \\ P^{*(n-1)} & n > 0 \end{cases} \tag{3}$$

The matrix W^* is defined in the following way

$$w^*_{i,j}(t) = \begin{cases} \frac{\beta(i,j)}{\sum_{d \in D(i,M(t))} \beta(i,d)} & i \in M(t) \wedge j \in M(t) \\ weight(e^*_{i,j}(t)) & e^*_{i,j}(t) \in E_{in}(t) \vee e^*_{i,j}(t) \in E_{out}(t) \\ 0 & otherwise \end{cases} \tag{4}$$

where $\beta : VxV \to \mathbb{R}$ is a function that indicates the strength and type of interaction between the pair of genes that takes as an input. The edge $e^*_{i,j}(t)$ is the pair of nodes (one virtual and one in the original metapathway) that is in the augmented metapathway.

3.1 Iteration algorithm

The iteration number could be seen as a quantified temporal time through the process of cell-to-cell communication and downstream propagation in the metapathways of the typed cell.

3.2 Things to define better

- Co-expression of genes in different cell-types profiles (the ligand-receptor pairs), this was already seen previously in the introduction but needed to be considered more carefully for the methodology, how to establish the interactions and the weight/probability of these interactions in CCC

- we do not consider weights for the edges between virtual nodes and genes.
- virtual nodes should have finer granularity, like virtual nodes for every LR pair.

For the first point(Establishing the interactions of cells via PPI or co-expression), there are many things to consider:

- Probability of interaction of protein units
- multi-subunit complexes(ligands complexes) with single receptors interactions, single interaction points from virtual nodes(representing the cell) and the receptor is an incomplete representation and too much simplistic.
- representing the interactions of ligands via hypergraphs to take into account the action of more ligands at the same time.

4 Notation table

Term	Mathematical definition	description
CT	$CT = \{t \mid \text{t is a cell type}\}$	set of cell types
t_s	$t_s \in CT$	cell type, the subscript means that it is seen as a source
t_d	$t_d \in CT$	cell type, the subscript means that it is seen as a destination
t	$t \in CT$	generic cell type
$M(t)$	$M : CT \to (G, E)$	metapathway associated with cell type t
$V^*(t)$	$V^*(t) = V(t) \cup V_{\text{in}}(t) \cup V_{\text{out}}(t)$	augmented metapathway nodes for cell type t
$E^*(t)$	$E^*(t) = E(t) \cup E_{\text{in}}(t) \cup E_{\text{out}}(t)$	augmented metapathway nodes for cell type t
$v_{t_s \to t}$	$v_{t_s \to t} \in V^*(t) \mid t_s \in CT, t \in CT$	virtual input node associated with type t
$v_{t_d \leftarrow t}$	$v_{t_d \leftarrow t} \in V^*(t) \mid t_d \in CT, t \in CT$	virtual output node associated with type t
$v^{(n)}_{t_s \to t}$	$v^{(n)}_{t_s \to t} : CT \times CT \times \mathbb{N} \to \mathbb{R}$	virtual input node value associated with virtual node $v_{t_s \to t}$ at the n-th time
$v^{(n)}_{t_d \leftarrow t}$	$v^{(n)}_{t_d \leftarrow t} : CT \times CT \times \mathbb{N} \to \mathbb{R}$	virtual output node value associated with virtual node $v_{t_d \leftarrow t}$ at the n-th time
co-expressed(g_i, g_j)	co-expressed : $V(t_i) \times V(t_j) \to \{TRUE, FALSE\}$	boolean-valued function that returns true if the two genes in the two metapathways are co-expressed (intercellular co-expression)
$V_{\text{in}}(t)$	$V_{\text{in}}(t) = \{v_{t_s \to t} \mid t_s \in CT \setminus t\}$	set of input nodes for the cell type t
$V_{\text{out}}(t)$	$V_{\text{out}}(t) = \{v_{t_s \leftarrow t} \mid t_s \in CT \setminus t\}$	set of output nodes for the cell type t
$Ligands(t)$	$Ligands(t) = \{v_i \mid v_i \in V(t), v_i \mapsto knownLigand\}$	set of all known ligands in the metapathway for type t
$Receptors(t)$	$Receptors(t) = \{v_i \mid v_i \in V(t), v_i \mapsto knownReceptor\}$	set of all known receptor in the metapathway for type t
$E_{\text{in}}(t)$	$E_{\text{in}}(t) = \{(v_{t_s \to t}, g_r, 1) \mid g_r \in Receptors(t), \exists g_l \in V(t_s) \mid g_l \in Ligands(t_s), \text{co-expressed}(g_l, g_r)\}$	set of input interactions from a virtual node to a node in the metapathway $M(t)$ for cell type t
$E_{\text{out}}(t)$	$E_{\text{out}}(t) = \{g_l, (v_{t_d \leftarrow t}, 1) \mid g_l \in Ligands(t), \exists g_r \in V(t_d) \mid g_r \in Receptors(t_d), \text{co-expressed}(g_r, g_l)\}$	set of output interactions from a node in the metapathway $M(t)$ to a virtual node for cell type t