Discovery of potential functional paths by integration of phospho-proteomics data in the PPI network using a RWR framework

Jérémie Perrin, Olivier Destaing, C. Brun

Discovery of potential functional paths by integration of phospho-proteomics data in the PPI network using a RWR framework

Jérémie Perrin[1], Olivier Destaing[2] and Christine Brun[1,3]

[1] Aix-Marseille University, INSERM, TAGC, Turing Centre for Living Systems, 163 Avenue de Luminy, Marseille 13009, France

[2] Institute for Advanced Biosciences, Centre de Recherche Université Grenoble Alpes, Inserm U 1209, CNRS UMR 5309, 38706 La Tronche, France

[3] CNRS, 31 Chemin Joseph Aiguier, Marseille 13009, France

Corresponding author: jeremie.perrin@univ-amu.fr

Abstract *Understanding how cellular signalling is flowing from the molecular to the cellular level is a key step to identify regulators of different diseases and revisit the development of new potential drug targets. For years, biological approaches of signalling did not allow to probe and control signalling at the sub-cellular level with enough accuracy in space and time to directly witness transfer of information in biological network. To analyze datasets where signaling is controlled spatio-temporally by optogenetic, we have developed a method which traverses the space of Random Walks with Restart (*RWR*) models, searching for the optimally biased walk in a given context. It will allow to integrate data of differentially phosphorylated proteins obtained from longitudinal phospho-proteomics assay, in response to two different mode of optogenetic activation of the kinase Src, in order to reconstruct potential functional paths in the Protein-Protein interaction (*PPI*) network.*

Keywords Random Walk with Restart, optimization, PPI network, path finding, signalling

1 Introduction

Activation of a single intracellular signaling element can induce a decision making event: different mode of activation of the same cell can have very different phenotypic responses. This suggests that some mechanism downstream of the stimulus drives the signalling processes into two signalling directions, inducing two different cellular responses. In an optogenetic engineered system, Kerjouan et al. [1] show such phenomenon at play. They successfully construct a functioning photo-activable version of the Src tyrosine kinase. By either being able to restrict the kinase movements to the membrane surface (2D diffusion) or being able to let it diffuse freely inside the cytosol (3D diffusion), they manage to activate the same level of Src in seemingly similar situations. They show that slightly modulating diffusion of these signals in the same site of action is sufficient to induce very distinct cellular phenotypes. In the case of membrane diffusion the cells exhibit lamellipodia, whereas they exhibit invadosome structures when the kinase is not restrained.

In order to understand the mechanisms at play, we need to be able to reconstruct the signal transduction after the activation of the Src kinase. Since our ability in monitoring the phosphorylation statuses of all proteins in a cell is fairly limited, the phospho-proteomics approaches are costly and their ability to be precisely quantitative is questionable, our hope is to be able to reconstruct some potential transduction paths from a coarse time-resolved phospho-proteomic assay, leveraging the information contained in the topology of the PPI network. We develop our method in order to decipher the events occurring between the time of the optogenetic Src activation and the cellular responses observed. A visual overview of the method is given in Figure 1.

2 Methods

To reconstruct potential functional paths, we observe the change in phosphorylation levels of a wide range of proteins following the activation of the optoSrc (OS-sensitive proteins). We use the RWR paradigm (see section 2.1) to analytically compute how much the Src kinase is able to influence the rest of the PPI network. The ordinary RWR makes a strong assumption on the possible interactions

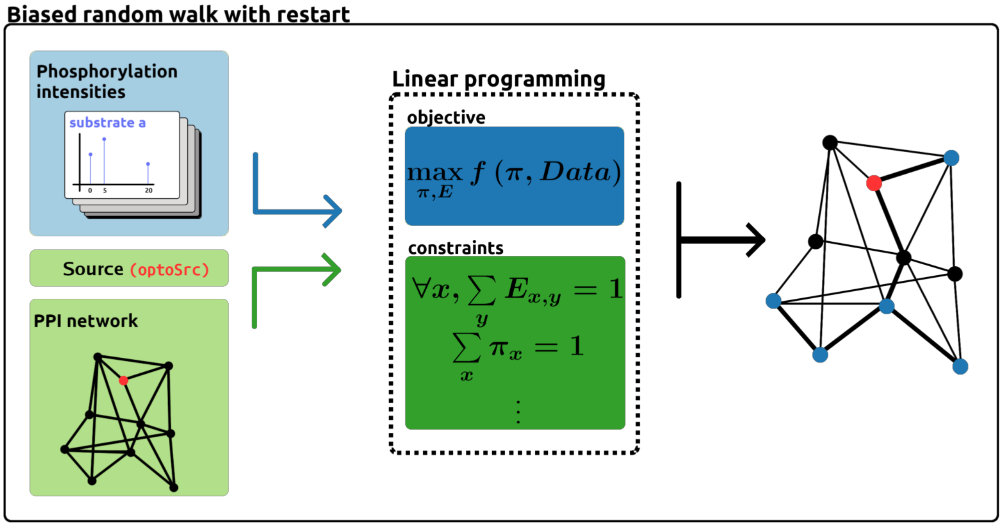
Biased random walk with restart
Phosphorylation intensities
substrate a
0
5
20
Source (optoSrc)
PPI network
Linear programming
objective
$\max_{\pi,E} f(\pi, Data)$
constraints
$\forall x, \sum_y E_{x,y} = 1$
$\sum_x \pi_x = 1$
$\vdots$

Fig. 1. *(left)* Inputs of the algorithm: phosphorylation intensities, topology of the PPI network and the protein of interest *(center)* Formulation of the RWR problem as a linear program *(right)* Output of the algorithm: the biases of the optimal RWR model which we will analyse in downstream workflow.

a protein might have at a given time : it supposes that a given protein has uniform probability of interacting with any of its partners, leading to a sphere of influence which is not context dependent. In our setting we have the information that, after a certain time, the activation of our initial kinase led to specific changes in phosphorylation levels. We will use this knowledge to guide our random walk and deduce some potential contextual affinities between protein partners. By exactly describing the space of RWR models using linear constraints on both the edge weights and the asymptotic distribution (see section 2.2), as was done in [2] for random walks, we are able to use cutting-edge optimization software [3] to find the mathematically optimal edge weights for a given objective function. The objective function we will be interested in is the quantitative matching of the RWR's asymptotic distribution to the experimental observation (see section 2.4). We therefore retrieve the edge weights corresponding to a stable distribution which matches best our observations of phosphorylation levels.

2.1 Random Walk with Restart

Given a graph $\mathcal{G} = (V, E)$, where V is the set of vertices (nodes) of the graph and E the set of edges, a set of initial nodes $\mathcal{I}$ with associated vector R ($\forall i \in I, R_i = \frac{1}{|\mathcal{I}|}$ and $\forall i \in V, i \notin \mathcal{I} \implies R_i = 0$). We can define A the adjacency matrix of the graph. We will start by defining the set of matrices we are interested in :

Definition 2.1 (Stochastic matrices with support A).

$$\mathcal{S}_A := \left\{ M \in M_n(\mathbb{R}) \quad s.t \quad \forall j, \sum_i M_{i,j} = 1 \; and \; \forall i, j, A_{i,j} = 0 \implies M_{j,i} = 0 \right\} \tag{1}$$

We define $\mathcal{S}_A$ the set of stochastic matrices which have A as a support. These matrices are stochastic because they represent a Markovian process : the random walk. This random walk is restricted to the edges of our underlying graph that is why the matrices should have the same support as A. This is a description of the space of parameters for our random walks.

For any matrix $W \in \mathcal{S}_A$, the Random Walk with Restart using the weights W is defined as the process:

$$P_{t+1} = \beta R + (1 - \beta) W P_t$$

Ergodic theory (using the Perron-Frobenius theorem) shows that such a process converges. We will call the asymptotic distribution of the process $\mathrm{RWR}_\beta(W)$. If $\Pi = \mathrm{RWR}_\beta(W)$ then it satisfies :

$$\Pi = \beta R + (1 - \beta) W \Pi \quad \text{and} \quad \Pi = \sum_{k=0}^{\infty} \beta (1 - \beta)^k W^k R$$

In most cases, when there is no information on the weights, W is chosen to be uniform at each nodes *i.e* $W = A^T D^{-1}$ where D is the diagonal matrix of degrees, we will be referring to this choice as the ordinary RWR.

The Random Walk with Restart (*i.e* Network Diffusion, Personalized PageRank) is a classical procedure which computes similarity scores between nodes in a graph. It is often used in network biology, and network science in general, to either : determine proximity between pairs of nodes in biological networks [4] and more complex networks [5], assess quality of general clustering methods [6] and to predict directionality in undirected PPI networks [7].

The PPI network constructed from Y2H high throughput experiments does not contain information about affinity of interactions nor context dependent competition between proteins. Although the RWR can leverage edge weights in weighted networks, due to the lack of information, the choice is often made to use uniform weights but it is a strong assumption. We have devised a procedure to optimize the weights in order to match some observations on the network. The choice of the network that we will actually use in the context of the Src signalling will not be discussed here but we can point out that it is one of the educated choices which has to be made in order to have the best possible predictions.

2.2 Optimally biased RWR

If we have some observations on a subset $\mathcal{O} \subset V$ of the vertices at two different time points $\{t_0, t_1\}$ (*i.e* we have a function $obs : \mathcal{O} \times \{t_0, t_1\} \to \mathbb{R}^+$). We then are interested in finding the optimal parameters which explain the change in the observations.

Definition 2.2 (Set of pairs of matrices with their associated asymptotic distribution). *To every matrix we can associate a unique asymptotic distribution under the biased random walk with restart with parameters ϵ and β.*

$$\mathbb{RWR}_{\epsilon,\beta} \coloneqq \{(S, RWR_\beta((1-\epsilon)W_0 + \epsilon S)), \quad S \in \mathcal{S}_A\} \tag{2}$$

Instead of searching the whole space of stochastic matrices which have A as a support $\mathcal{S}_A$ (see Equation 1), which could give very unpractical results because it would not take into account diffusion at all, we will search the space of parameters around an *a priori* set of parameters W_0. The choices of W_0 could be one of the following: uniform probabilities, maximum entropy probabilities, optimized probabilities to match the initial observation (the actual choice will not be discussed here but is a crucial point to be studied in the future). To summarize, we will be considering, for a given ϵ, all matrices of the form $(1-\epsilon)W_0 + \epsilon S$ for S in $\mathcal{S}_A$, those are the biased random walks around W_0 (see Equation 2). Out of all of these matrices we will try to find the one which best explains the changes in the observations. If we have a function f to compare the asymptotic distribution of the RWR to the observations then we are interested in :

$$S^* \in \underset{S,\Pi \in \mathbb{RWR}_{\epsilon,\beta}}{\operatorname{argmin}} f(\Pi, obs) \tag{3}$$

Remark 2.3. *In our case, the observations will be the phosphorylation intensities measured for each protein for which we actually have a value, before and after Src activation. We will discuss the actual choice of f in section 2.4, but first let us describe the space of all possible RWRs.*

Remark 2.4. *These optimization problems are hard in general, except when the problem can be formulated in certain ways. This is the theory of convex optimization. We will not address the mathematics underlying the optimization procedures but we will show that our problem can be formulated as a Linear Program. Linear Programs are a kind of well studied optimization problems for which we have good enough optimizers for problems of the size we are interested in.*

2.3 Specifying the convex search space

We want to describe the set $\mathbb{RWR}_{\epsilon,\beta}$, the stochastic matrices with support matrix A as well as their associated asymptotic distribution for the Random Walk with Restart. We will give the linear constraints which describe this set as a subspace of $\mathbb{R}^{|V|^2} \times \mathbb{R}^{|V|}$.

Let us start with $\mathcal{S}_A$:

$$S \in \mathcal{S}_A \iff \forall j, \quad \sum_{i=0}^{|V|} S_{ij} = 1 \text{ and } \forall ij, \quad 0 \leq S_{ij} \leq A_{ji} \tag{4}$$

Now let us describe $\mathrm{RWR}_\beta((1-\epsilon)W_0 + \epsilon S)$:

$$\begin{aligned}\Pi = \mathrm{RWR}_\beta((1-\epsilon)W_0 + \epsilon S) \iff \quad & \sum_j \Pi_j = 1 & (5)\\ \text{and} \quad & \forall j, \quad 0 \leq \Pi_j & (6)\\ \text{and} \quad & \Pi = \beta R + (1-\epsilon)(1-\beta)W_0\Pi + \epsilon(1-\beta)S\Pi & (7)\end{aligned}$$

Here Equation 7 is not linear, the problem is solved by a simple change of variables $E \coloneqq S \times \mathrm{diag}(\Pi)$ and rewriting the equations subsequently (similarly to what is done in [2]):

$$\forall j, \quad \sum_{i=0}^{|V|} E_{ij} = \Pi_j \text{ and } \forall ij, \quad 0 \leq E_{ij} \leq A_{ji} \tag{4bis}$$

$$\Pi = \beta R + (1-\epsilon)(1-\beta)W_0\Pi + \epsilon(1-\beta)E \tag{7bis}$$

2.4 Objective function

Let us discuss the choice of f, in order for the problem to be a Linear Program, we need f to be linear in both E and Π. Since our goal is to match the observations, it actually will not depend on E (although we could later add some regularization term to our objective which could depend on E). In fact we already hinted that way in Equation 3, by not having E be an argument of f.

Our goal is to find the edge weights which bias the random walk's asymptotic distribution towards the activated nodes and away from the inactivated nodes. Thus our first approach was to consider the sets of nodes $\mathcal{O}^+ \coloneqq \{i \in \mathcal{O}, \quad obs(t_1) - obs(t_0) > 0\}$ and $\mathcal{O}^- \coloneqq \{i \in \mathcal{O}, \quad obs(t_1) - obs(t_0) < 0\}$. Trying to maximize probabilities of ending on nodes in $\mathcal{O}^+$ which have a positive variation of the observed quantity and minimize probabilities of ending on nodes in $\mathcal{O}^-$ which have negative variation of observed quantity :

$$f_1(\Pi, obs) = -\left(\sum_{i\in\mathcal{O}^+} \Pi_i - \sum_{i\in\mathcal{O}^-} \Pi_i\right)$$

By not taking into account the actual structure of RWR (*i.e* correlation between distance from the source and the probability of ending on the node), this tends to try and optimize the probabilities on nodes that are closest to the initial nodes. We solved this issue by looking at the fold change in probability from the initial condition $\mathrm{RWR}(W_0)$:

$$f_2(\Pi, obs) = -\left(\sum_{i\in\mathcal{O}^+} \frac{\Pi_i}{\mathrm{RWR}(W_0)_i} - \sum_{i\in\mathcal{O}^-} \frac{\Pi_i}{\mathrm{RWR}(W_0)_i}\right)$$

Now this was satisfactory in practice, but it didn't leverage the whole information of the observations: we only considered the trend of the observed quantity, not the actual values. We would rather be matching the actual fold change in the observed quantity :

$$f_3(\Pi, obs) = \sum_{i\in\mathcal{O}} \left|\frac{\Pi_i}{\mathrm{RWR}(W_0)_i} - \frac{obs(i,t_1)}{obs(i,t_0)}\right|$$

The absolute values are not linear, but there exists a trick which introduces new variables in the Linear Program to transform absolute values in the objective function into constraints :

$$\begin{aligned} \text{Minimizing } f_3 \iff & \quad \text{Minimizing } f_4 = \sum_{i \in \mathcal{O}} X_i \\ & \text{and} \quad X_i \geq \frac{\Pi_i}{\mathrm{RWR}(W_0)_i} - \frac{obs(i,t_1)}{obs(i,t_0)} \\ & \text{and} \quad X_i \geq -\left(\frac{\Pi_i}{\mathrm{RWR}(W_0)_i} - \frac{obs(i,t_1)}{obs(i,t_0)}\right) \end{aligned}$$

2.5 Solving

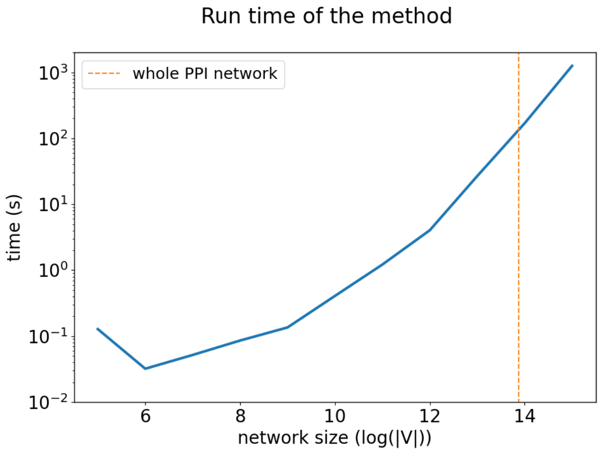

Fig. 2. Run times of the method on synthetic data. We have generated networks of different sizes, using Watts–Strogatz model. We have kept the ratio of number of edges and number of vertices fixed in all generated graphs. We consistently used a single source and $\frac{|V|}{30}$ target nodes.

Now that we have specified our problem as a Linear Program in a satisfactory way, we can compute the optimal solution *i.e* the optimal deviation from the initial condition to match our observations. If we come back to our biological question, we have a description of those interactions which are potentially favored or unfavored, in the context of our observations.

We are using the gurobi optimizer in python, through the gurobipy package. We have an academic license, but the software is proprietary. In order to support open source ecosystems, we might use the CLP software [8] in the future.

Remark 2.5 (Scalability). *Scalability does not seem to be an issue, see Figure 2, although restricting the network to functional modules will definitely run faster than on the whole PPI network.*

3 Results

As for now, we have tested our method on synthetic data. A visual representation of the output of the method on a 6 × 6 grid with single source and two positive targets is given in Figure 3.

From the output of our method we are capable of extracting paths from sources to targets, either from the probability landscape (Figure 3.A) or from the bias of our optimal RWR (Figure 3.B). In the actual workflow on the Src activation problem, we will develop some further downstream analysis to study the paths. In the Src context we will actually get a couple of paths we will need to compare in order to determine proteins/interactions which are potentially responsible for the decision making. We will also discuss validation of the method on gold standard data in section 4.

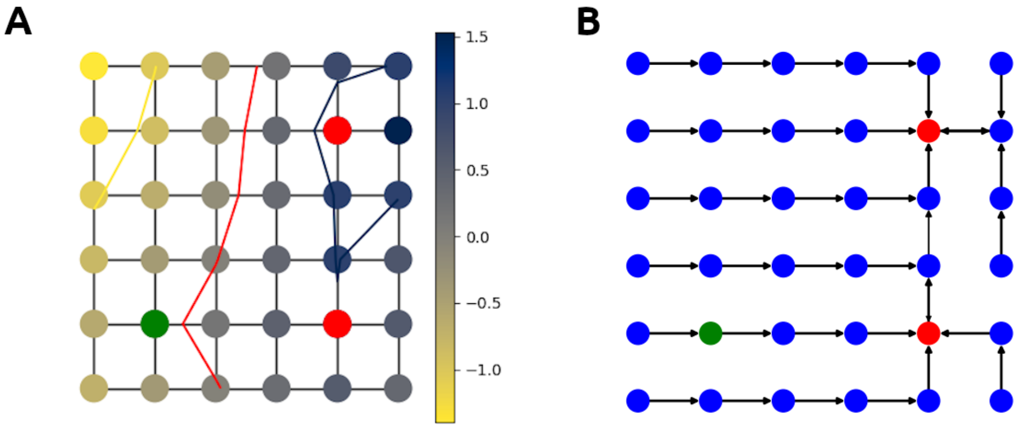

Fig. 3. **(A & B)** Green node is the source node, both red targets have a positive variation in the observed quantity **(A)** Probability landscape as a $\log_2$ fold-change between RWR(W_0) and RWR($(1-\epsilon)W_0 + \epsilon S^*$) **(B)** Representation of S^* as a directed graph, we visualize the biases *i.e* the directions favored in order to optimize the probabilities on the target nodes.

4 Discussion

We have presented our method, some technical details and some choices we have made based on our understanding of both the mathematical tool we are using and the biological context in which we want to apply the method. This being a work in progress, a substantial amount of leeway remains. One could question the biological relevance of the paths we find, and would be legitimate in doing so. We do not pretend to be providing a model of signalling, the potential functional path we are discovering will help us guide new experiments and question the roles of specific proteins, but should not be seen as predicted signalling pathways. In order to reinforce our trust that there is biological significance in the paths we discover, we will have to validate our method and adapt it to integrate other sources of biological information.

There are a couple of instances where we can, and will, try to validate our approach. The first case is in trying to reproduce directionality information as was done in [7]: we can try to predict the direction of interactions and confront our predictions to the ground truth since some interactions are known to be directed. Another idea is in trying to retrieve signalling pathways, from an obfuscated version of said pathways. Indeed, we could consider a signalling pathway, forget the directionality of the interactions, add some partner proteins which are not part of the pathway (obfuscating the original pathway) and then see if our method is capable of retrieving some parts of the pathway. Seeing how the method fares in controllable settings, will define how relevant we consider the paths discovered *de novo*.

A criticism regarding the interpretability of the RWR in the context of the PPI network comes from the fact that the PPI is constituted of binary interactions which are tested in conditions very different from the condition in which the interactions actually occur. The ”real” network representation of the protein interaction at a given time in a given cell has to differ from the PPI network (some interactions might be context-dependant). This observation drives our work in the direction of integrating other sources of information into the RWR paradigm. We are currently working on different ways of integrating orthogonal data to the Random Walk with Restart (proteomics, functional annotation). We have already mentioned that the choice of the underlying network is crucial, we have already built sub-networks of the PPI from proteins extracted from the literature, and are thinking about integrating data from Phosphosite+ [9]. We have developed our method in order to be able to integrate cause-to-effect contextual information, adding even more biological information should bring more specific results.