Integrating Protein Localization with Automated Signaling Pathway Reconstruction

Ibrahim Youssef
Biology Department
Reed College
Portland, USA
youssefi@reed.edu

Jeffrey Law
Department of Computer Science
Virginia Tech
Blacksburg, USA
jeffl@vt.edu

Anna Ritz
Biology Department
Reed College
Portland, USA
aritz@reed.edu

***Abstract*—Understanding cellular responses via signal transduction is a core focus in systems biology. Tools to automatically reconstruct signaling pathways from protein-protein interactions (PPIs) can help biologists generate testable hypotheses about signaling. However, automatic reconstruction of signaling pathways suffers from coarsely-weighted interactions, leading to many equally good candidates. Further, some reconstructions are biologically misleading due to ignoring protein localization information. We propose *LocPL*, a method to improve the automatic reconstruction of signaling pathways from PPIs by incorporating information about protein localization in the reconstructions. The method relies on a dynamic program to ensure that the proteins in a reconstruction are localized in cellular compartments involved in signaling transduction and that the interactions are consistent with signaling from the membrane to the nucleus. *LocPL* produces more accurate and biologically meaningful reconstructions on a versatile set of signaling pathways. The code, including a newly-released interactome $PLNet_2$, is available at https://github.com/annaritz/localized-pathlinker.**

***Index Terms*—Signal transduction, Biological networks, Data integration, Pathways, Protein-protein interaction, Protein Localization**

I. Introduction

A fundamental goal of molecular systems biology is to understand how individual proteins and their interactions may contribute to a larger cellular response. Repositories of experimentally derived human protein-protein interaction (PPI) information [1]–[4] have been critical for studying the topology of signaling pathways. These datasets have enabled the development of methods that aim to link extracellular signals to downstream cellular responses, which are characterized as signaling pathways. Pathway analysis methods conceptualize the interaction information as a graph, or an *interactome*, where edges connect proteins that are known to interact experimentally. Here, we will focus on methods that identify static networks to characterize the potential topology of human signaling pathways. Approaches for identifying relevant subnetworks have drawn on different graph theoretic methods, including shortest paths [5]–[8], Steiner trees and related formulations [9], [10], network flow [11], [12] and random walk approaches [13], [14].

PathLinker is a recent pathway reconstruction approach that returns ranked paths for a specific human signaling pathway of interest [7]. Given a weighted interactome, a set of known receptors, and a set of known transcriptional regulators (TRs), PathLinker returns the *k*-shortest paths from any receptor to any transcriptional regulator. PathLinker was shown to have superior performance over other graph-based methods such as Steiner trees, network flow, and random walks.

Pathway Reconstruction Challenges. Despite PathLinker's success, the problem of identifying accurate pathway reconstructions remains challenging. PathLinker paths are prioritized by their reconstruction scores, and the collection of these paths constitute a *pathway reconstruction*. We assessed PathLinker reconstructions for four well-studied and diverse signaling pathways: Wnt, Interleukin-7 (IL-7), $\alpha 6\beta 4$ Integrin and Epidermal Growth Factor Receptor (EGFR1). Careful analysis of the ranked paths across these pathways revealed two main challenges in pathway reconstruction.

First, we found that many PathLinker paths have identical reconstruction scores. For example, about 52% of the paths in the Wnt reconstruction had the same score. This feature was not unique to Wnt; 60%, 82.6%, and 48.2% of the paths were tied in the IL-7, $\alpha 6\beta 4$ Integrin, and EGFR1 pathways, respectively. Strikingly, even the top-ranked paths in the reconstructions were often tied (top 38 paths in Wnt, top 43 paths in IL-7, top 57 paths in $\alpha 6\beta 4$ Integrin, and top 330 paths in EGFR1). We found that the tied paths were a result of many interactions with identical weights in the underlying interactome (Fig. 1). For example, in the PathLinker interactome ($PLNet_1$), nearly 68% of the interactions have only two distinct weight values. The coarse interaction weighting is also apparent in the HIPPIE network [2], where 55% of the interactions share the same edge weight (Fig. 1).

Second, we noted that paths in the reconstructions contained a mix of pathway-specific signaling interactions relevant to the

AR is supported by the National Science Foundation (1750981) and the M. J. Murdock Charitable Trust (92015288:MNL:2/25/2016). JL is partially supported by the NIH-NIGMS (R01-GM095955-01) and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the Army Research Office (ARO) under cooperative Agreement Number W911NF-17-2-0105. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, ARO, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

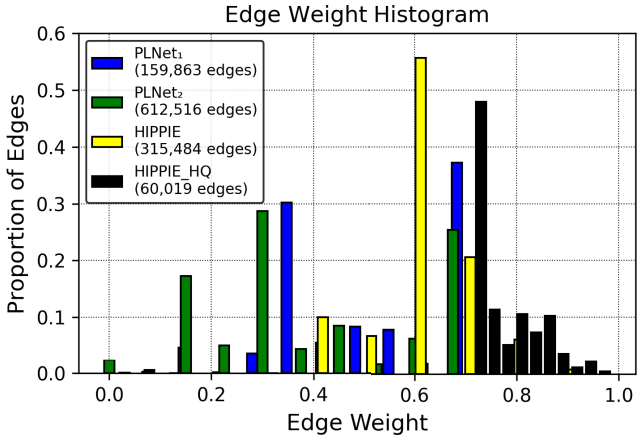

Fig. 1. Proportion of edges with identical edge weights in the PathLinker and HIPPIE interactomes. *PLNet*$_1$ is the PathLinker interactome [7], while *PLNet*$_2$ is the interactome used in this work. The HIPPIE High Quality (HIPPIE_HQ) interactome includes all HIPPIE edges with a weight ≥ 0.73 [2].

pathway under study (*positive interactions*) and non-pathway interactions (we will call them *negative interactions*, though they may very well be signaling interactions relevant to other pathways or pathway-specific interactions that have not been annotated yet). Paths are rarely comprised solely of positive interactions: in all four pathway reconstructions, over 95% of the paths that include at least one positive interaction also contain a negative interaction. PathLinker does not consider protein localization in the pathway reconstructions, so interactions within the same path may be unrealistic in terms of compartment co-localization. Given the first challenge of coarse interaction weights, incorporating biologically-motivated path constraints could be useful for breaking tied path scores (Fig. 2, Left).

To overcome the challenges described above, we sought to incorporate an independent data type into the pathway reconstruction problem. Many methods have integrated gene expression data to uncover signaling pathways relevant to a particular condition or disease [9], [11], [14] ; however incorporating gene expression changes may alter the "canonical" pathways that reconstruction methods such as PathLinker aim to predict. Instead, we make use of information about a protein's localization within the cell to constrain the paths in a reconstruction.

Contributions. We propose *LocPL*, an extended version of PathLinker that reconstructs pathways by incorporating information about cellular localization in two ways. First, *LocPL* uses localization information to discard likely false positive interactions from the interactome before running PathLinker, improving its specificity. Second, *LocPL* incorporates the localization information in a dynamic programming scheme to identify spatially-coherent paths and re-prioritize tied paths. Previous work (based on a node-coloring scheme) constrained paths by requiring that the path is partitioned into segments, where each segment consisted of nodes belonging to a certain category of proteins [6]. Our approach, on the other hand, computes the most likely compartments for each node in the path and outputs a score that reflects this likelihood. The path length is also considered an input for the node-coloring method, while *LocPL* lifts this constraint.

We show that paths with larger proportions of signaling interactions will be promoted higher in the *LocPL* k-shortest paths list.We compare the *LocPL* pathway reconstructions to those from PathLinker and illustrate the changes that result from constraining the paths by cellular compartments. We provide a new interactome, *PLNet*$_2$, which quadruples the number of interactions compared to the PathLinker interactome. In addition to performing a global performance assessment of paths, we present a local measure to assess path quality individually. Visual inspection of the top 100 paths in the Wnt, IL-7, $\alpha 6\beta 4$ Integrin, and EGFR1 pathway reconstructions reveal that the spatially-coherent approach changes the reconstruction topology, in some cases removing paths that lead to activation of other pathways. This work demonstrates that incorporating protein localization information into signaling pathway reconstruction improves predictions that are necessary for appropriate hypothesis generation.

II. Methods

We first present an overview of *LocPL*, which uses information from a protein localization database to refine pathway reconstructions. After describing the model used for signaling flow, we present a dynamic program for computing scores that reflect a path's consistency with the model of signaling. Finally, we describe the datasets and the means of assessing pathway reconstruction performance.

A. LocPL: Localized PathLinker

Signaling pathway analysis methods typically take an interactome as input, represented as a graph $G = (V, E)$ where the nodes V are proteins and the edges E are PPIs. In the case of *LocPL*, the graph is directed, each edge $(u, v) \in E$ has a weight $w_{uv} \in [0, 1]$, and every interaction is predicted to occur within some cellular compartment according to a database of protein localization information. *LocPL*'s core method is a k-shortest path algorithm previously described as PathLinker [7]. Given a directed, weighted interactome G, a set R of receptors and a set T of transcriptional regulators (TRs) for a pathway of interest, and a number of paths k, PathLinker outputs a ranked list of the k shortest paths, $\mathcal{P} = \langle P_1, P_2, \ldots, P_k \rangle$, where a path $P_i = (v_1, v_2, \ldots, v_m)$ is comprised of m nodes that begin at a receptor ($v_1 \in R$) and ends at a TR ($v_m \in T$). Each path P_i is ranked by the product of its edge weights (its *reconstruction score* r_i), and $r_i \geq r_{i+1}$ for every i.

After running PathLinker on the interactome, *LocPL* breaks ties in the candidate list of paths $\mathcal{P}$ by considering a model of signaling flow based on cellular compartments. For each path P_i, a dynamic program identifies the *signaling score* s_i of the most likely series of compartments for each node that is consistent with the signaling flow model. After this step, each path P_i will have two scores: a reconstruction score r_i computed by PathLinker and a signaling score s_i computed by the dynamic program. The signaling score is used to re-prioritize

the tied reconstruction scores by partitioning the paths into ties (e.g. all paths with the same reconstruction score) and reordering the paths within each group in decreasing order of the signaling score.

B. Localized Protein-Protein Interactions from ComPPI

ComPPI is a database that predicts cellular compartments for human proteins and PPIs [15]. For each protein, ComPPI computes *localization scores* describing the likelihood of a protein to be found in one of the major six subcellular compartments: (i) extracellular fluid, (ii) cell membrane, (iii) cytosol, (iv) nucleus, (v) secretory pathway (e.g. transport vesicles), and (vi) mitochondria. ComPPI uses three types of information to infer the localization scores: experimental verification, computational prediction, and unknown sources, resulting in high, medium, and low localization scores, respectively. The *interaction score*, computed by ComPPI from localization scores of the participating proteins, represents the probability that an interaction takes place inside the cell.

LocPL uses the ComPPI database to restrict the interactions of the interactome by removing edges with an interaction score of zero – these interactions could take place from a biophysical perspective, but are less likely to occur within the cell due to the predicted protein localization. After this filtration step, all edges in the interactome have a non-zero probabilistic score aggregated across all cellular compartments. For subsequent steps of *LocPL*, we use the ComPPI localization scores that reflect individual proteins in specific cellular compartments.

C. Signaling Flow Structure and Assumptions

We first state some assumptions about the pathways we aim to reconstruct, though the model is flexible and the assumptions may be customized for a specific pathway of interest. First, we only consider intracellular signaling that begins with activation of a membrane-bound protein receptor and is transmitted to a DNA-binding transcription factor through PPIs within the cytosol. Second, we focus on three cellular compartments: a combination of extracellular fluid and cell membrane (*ExtMem*), which represents where a receptor may be located, *Cytosol*, and *Nucleus*. Third, we assume a fixed unidirectional signaling flow that follows a structure composed of compartmental layers within the cell, from *ExtMem* through *Cytosol* to *Nucleus*. Fourth, multiple interactions may occur within the same compartmental layer (e.g. multiple interactions may occur within *Cytosol*). These assumptions impose an ordering on the compartments that must be visited, which we will use in breaking tied paths. Fig. 2. (Right) illustrates these assumptions with four different paths as examples of valid and invalid paths/interactions. Paths **a** and **b** are valid; however, path **c** is not valid because signaling goes directly from the cellular membrane to the nucleus and path **d** has one invalid interaction because signaling goes in a direction against the assumed signaling flow.

We acknowledge that these assumption may not hold for many pathways. For example, some pathways are initiated via

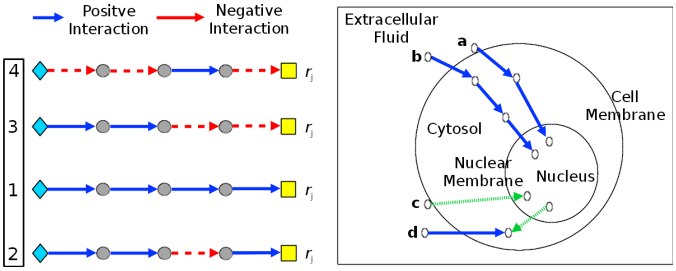

Fig. 2. **Left.** Illustration of four PathLinker paths from receptors (diamonds) to transcriptional regulators (yellow boxes) that all have the same reconstruction score r_j. Blue solid edges represent true positive interactions, and red dashed edges represent false positives. The goal of breaking ties is to re-rank the tied paths so paths with more positives are ranked higher (black box). **Right.** Simplified model diagram for the signaling flow structure. Blue solid edges represent valid interactions, while the green dotted edges represent invalid interactions.

nuclear receptors, and would be missed based on our assumption that signaling begins at receptors at the cell membrane. We also do not consider other compartments beyond *ExtMem*, *Cytosol*, and *Nucleus* in our model, while the mitochondria and secretory vesicles play an important role in some signaling pathways. The model of signaling flow may be customized to each pathway under study, and *a priori* information about the structure of signaling flow may further improve *LocPL* predictions.

D. Dynamic Program for Path-Based Signaling Scores

Given a path $P = (v_1, v_2, \ldots, v_m)$ that connects m proteins, our goal is to find a selection of compartments that maximize the path signaling score (by sum of log-transformed localization scores) while respecting the signaling flow structure outlined in Section II-C. For each protein $v \in V$ that contains localization information, we use $\ell_v^{ext}, \ell_v^{cyt}$, and ℓ_v^{nuc} to denote the scores of *ExtMem*, *Cytosol*, and *Nucleus* respectively. We log-transform these scores to be localization costs, that is, $\ell_v^c = -\log \ell_v^c$ for each protein v and each cellular compartment c (either *ExtMem*, *Cytosol*, or *Nucleus*). Let $s(v_j, c)$ be the optimal score of the path up to node $v_j \in P$, where v_j is in compartment c. The optimal signaling score of the path must end in the nucleus, which we denote by $s(v_m, nuc)$. The score can be computed as:

$$s(v_m, nuc) = \min\left[s(v_{m-1}, cyt), s(v_{m-1}, nuc)\right] + \ell_{v_m}^{nuc}$$

This recurrence states that the largest score of the entire path (to v_m) ending in the nucleus is the sum of the localization score of protein v_m in the nucleus and the maximum of *either* (a) the largest score of the path up to v_{m-1} in the nucleus, or (b) the largest score of the path up to v_{m-1} in the cytosol. This formula is consistent with our assumptions that (1) the path must end in the nucleus, and (2) the last interaction must either be in the nucleus or must connect a protein in the cytosol to a protein in the nucleus. In general, at node v_j,

$j = 2, 3, \ldots, (m-1)$, the set of equations for the scores are:

$$s(v_j, ext) = s(v_{j-1}, ext) + \ell_{v_j}^{ext}$$
$$s(v_j, cyt) = \min\left[s(v_{j-1}, ext), s(v_{j-1}, cyt)\right] + \ell_{v_j}^{cyt}$$
$$s(v_j, nuc) = \min\left[s(v_{j-1}, cyt), s(v_{j-1}, nuc)\right] + \ell_{v_j}^{nuc}.$$

Note that we can only reach a protein in *ExtMem* from another protein in *ExtMem*, we can reach a protein in *Cytosol* from another protein in either *ExtMem* or *Cytosol*, and we can reach a protein in *Nucleus* from another one in either *Cytosol* or *Nucleus*.

To ensure that the path starts with the cellular compartment *ExtMem*, the base case for these recurrence relations are:

$$s(v_1, ext) = \ell_{v_1}^{ext},\ s(v_1, cyt) = \infty, \text{ and } s(v_1, nuc) = \infty.$$

These recurrence relations can be efficiently calculated using a dynamic program, filling an $m \times 3$ table denoting the number of nodes (m) by the three compartments. The final score taken will be $s(v_m, nuc)$ since we require the path to terminate in the nucleus.

E. Interactomes and Pathways

a) PLNet$_2$ Interactome: We built *PLNet*$_2$ from both physical molecular interaction data (BioGrid, DIP, InnateDB, IntAct, MINT, PhosphositePlus) and annotated signaling pathway databases (KEGG, NetPath, and SPIKE) [3], [4], [16]–[18]. *PLNet*$_2$ contains 17,168 nodes, 40,016 directed regulatory interactions, and 286,250 bidirected physical interactions, totaling 612,516 directed edges. We assigned interaction direction based on evidence of a directed enzymatic reaction (e.g., phosphorylation, dephosphorylation, ubiquitination) from any of the source databases. Each interaction was supported by one or more types of experimental evidence (e.g. yeast two hybrid or co-immunoprecipitation) that are available as evidence codes from the data sources, and/or the name of the pathway database it is from. Edges were weighted using an evidence-based Bayesian approach that assigns higher confidence to an experiment type/pathway database if it identifies interacting proteins that participate in the same biological process [11]. We chose the GO term "regulation of signal transduction" to build a set of positive interactions that are likely related to signaling; this term includes "signal transduction" as a child GO term. Positives are edges whose nodes are both annotated with this term, and negatives are randomly selected edges whose nodes are not co-annotated to the term. We chose $|N| = 10 \times |P|$ negative edges. To lessen the influence of very highly-weighted edges, we applied a ceiling of 0.75 to all weights [11].

b) Ground Truth Pathways: We considered the $\alpha 6\beta 4$ Integrin, EGFR1, IL3, IL6, IL-7, Wnt, RANKL, and TGF_Beta_Receptor pathways from the NetPath database [16] as our ground truth. We excluded other pathways such as Notch since its receptors have intracellular domains that are also TRs, violating the assumptions outlined in the model of signaling flow. Receptors and TRs are automatically detected for each of the eight pathways from lists of 2,124 human receptors and 2,286 human TRs compiled from the literature; see [7] for more details.

F. Global and Path-Based Assessment

We assess the performance of *LocPL* compared to PathLinker (*PL*) using two methods that evaluate global and local features of the ranked paths.

a) Precision-recall (PR) curves: Given a ranked list of paths (e.g. returned by *LocPL* or *PL*), we order each interaction by the index of the path in which it first appears. We compute precision and recall for this ranked list using the NetPath interactions as positives and a sampled set of negative interactions that are 50 times the size of the positive set. We also computed the aggregated precision and recall for paths from all the eight pathway reconstructions (aggregate pathways) using *PLNet*$_2$.

b) Path-based assessment: The PR curves provide a global quantitative assessment across all the k paths in a reconstruction, showing how quickly (in terms of k) the technique can discover new positive edges. However, this approach considers a positive only once, i.e., the first times it appears in a path. Thus, this global measure fails to characterize each path individually in terms of the number of positives contained in that path. Hence, we introduce a simple way to "locally" assess paths by computing the within-path percentage of true positive edges, denoted as *PosFrac*. Since we compute this metric value independently for each path, it does not matter if a positive interaction is detected earlier in another path. We computed the moving average of *PosFrac* values, using non-overlapping intervals of 100 paths each.

c) Statistical significance: For each assessment method, we use the Mann-Whitney U (MWU) statistical test for unpaired samples to estimate whether the difference between *PL* and *LocPL* results is statistically significant. The inputs to the MWU test for the path-based case are the *PosFrac* values of *PL* and *LocPL*. The global assessment is based on two concurrent values: precision and recall. These two quantities are related, so we use their harmonic mean (F_1 score) to get a single value summarizing the pair. We use the F_1 score values of *PL* and *LocPL* as the inputs to the MWU statistical test. We acknowledge that *PosFrac*, precision and recall are not purely independent between the two methods, so there is some dependence introduced in the MWU tests.

III. Results

A. Combining Interactomes with Localization Information

Approximately 80% of the proteins in *PLNet*$_2$ have localization information, producing an interactome with about 44% of the edges (Table I). Focusing on four pathways of interest ($\alpha 6\beta 4$ Integrin, EGFR1, IL-7, and Wnt) for space consideration, 50% of the edges have localization information, but 27 of 29 receptors and all 53 TRs have localization information and remain in the network. After filtering *PLNet*$_2$ using ComPPI, 93% of the proteins have a non-zero *ExtMem* localization score, 73% have a non-zero *Cytosol* localization score, and 61% have a non-zero *Nucleus* localization score.

TABLE I
NUMBER OF PROTEINS AND INTERACTIONS IN $PLNet_2$.

Interactome	Complete Interactome		Interactome ∩ ComPPI	
	Nodes	Edges	Nodes	Edges
$PLNet_2$	17,168	612,516	13,681	267,403

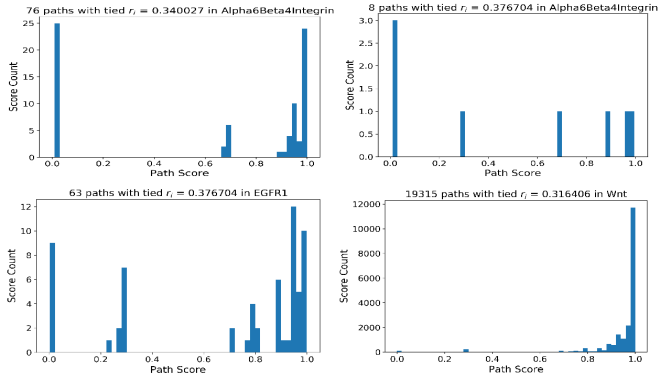

Fig. 3. Histogram of signaling scores s_i for four examples of paths with tied reconstruction score r_i (indicated in the title).

Most of the proteins have multiple non-zero localization scores for different compartments, though more proteins with a single non-zero localization score appear in the *Nucleus* than other compartments.

Applying PathLinker to the ComPPI-filtered interactome partially mitigates the problem of tied paths, but many ties remain. For example, after running PathLinker on the $\alpha 6\beta 4$ Integrin pathway with the full $PLNet_2$ interactome, there were 82 groups of paths where each group shared the same reconstruction score. This number was reduced to 38 groups when running PathLinker on the filtered $PLNet_2$ interactome. However, ties still dominate the reconstruction scores; thus the need for an approach to break these ties and re-prioritize paths in a biologically relevant way is still imperative.

B. Assessment of Pathway Reconstructions

We applied PathLinker (*PL*) and *LocPL* to signaling pathways from the NetPath database to the $PLNet_2$ as described in Section II-E. We computed $k = 20,000$ paths for each approach, similar to the original publication [7]. Paths that have the same reconstruction score differ substantially in their signaling scores computed by the dynamic program (Fig. 3).

a) Precision and Recall: Fig. 4 shows the PR curves used to globally assess *PL* and *LocPL* for four signaling pathways: $\alpha 6\beta 4$ Integrin, EGFR1, IL-7, and Wnt. *LocPL* generally outperforms *PL* in terms of precision and recall, where the precision of *LocPL* is greater than *PL* at nearly all values of recall. Moreover, *LocPL* usually detects higher proportions of positives than *PL* as reflected in the larger recall values for *LocPL*.

For every value of precision and recall, we plot the harmonic mean (F_1 score) of the two values for *LocPL* and *PL* (Fig. 5). The F_1 curve for *LocPL* is noticeably higher than that of *PL* for $\alpha 6\beta 4$ Integrin, EGFR1, and Wnt pathways (MWU test p-value ≤ 0.0044). However, it is slightly higher than that of

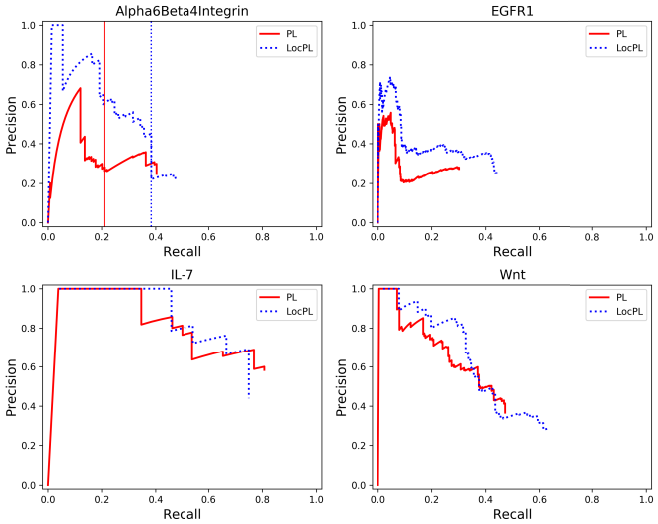

Fig. 4. Precision and recall curves of pathway reconstructions from PathLinker (*PL*) and *LocPL* on four NetPath signaling pathways. The dashed vertical lines of the $\alpha 6\beta 4$ Integrin pathway plot represent the recall value after reconstructing 5,000 paths.

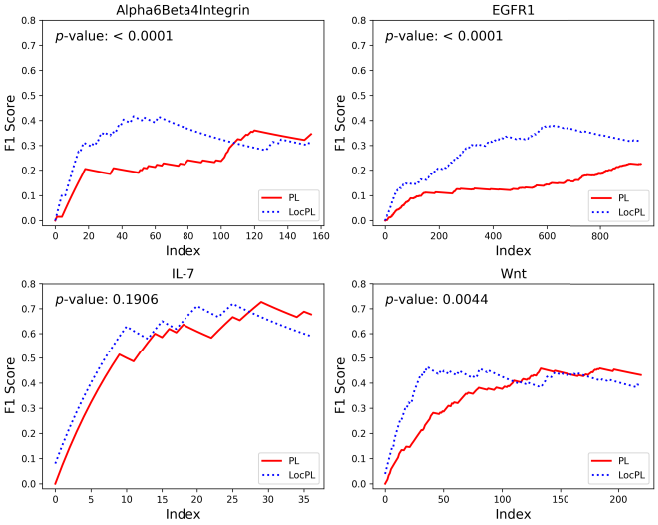

Fig. 5. F_1 scores for the individual NetPath pathways. The p-value is for the MWU test (alternative: *LocPL* > *PL*).

PL for IL-7, making this difference statistically insignificant (MWU test p-value of 0.1906).

b) Path-based Assessment: In addition to the global assessment, we are interested in the quality of subsets of paths. Plotting *PosFrac* of non-overlapping windows of 100 reveals subsets of paths that are enriched for positive interactions in the four pathway reconstructions (Fig. 6).[1] In the IL-7 pathway reconstruction, paths produced by *LocPL* tend to contain more positive signaling edges than those obtained by *PL* over all the 20,000 paths. *PosFrac* is almost consistent for *LocPL* and, despite some spikes (of different widths) for *PL*,

[1]Note that *PosFrac* considers all negative interactions for each path, unlike the PR curves in Fig. 4 that subsample the negative set of interactions. Thus, the *PosFrac* values will be smaller than one would expect based on the PR curves.

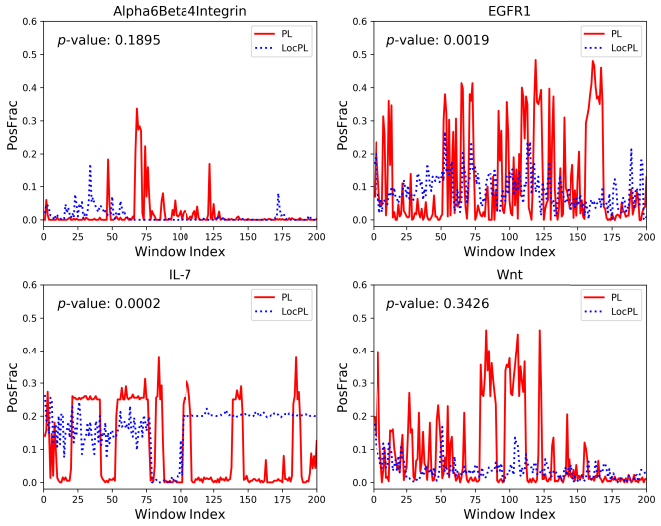

Fig. 6. Path-based performance of *PL* and *LocPL* on four NetPath signaling pathways. *PosFrac* is the percentage of positives averaged across non-overlapping windows of 100 paths. The p-value is for the MWU test (alternative: *LocPL* $>$ *PL*).

PosFrac for *LocPL* dominates the graph. In the IL-7 pathway reconstruction, this distinction is significant (one-tailed MWU test, Fig. 6). *LocPL* is also significantly better than *PL* for the EGFR1 pathway, which is due to *LocPL*'s low variance in *PosFrac* values, even though *PL* has more "peaks" of large *PosFrac* values.

The situation is different for the $\alpha 6\beta 4$ Integrin and Wnt pathways, where *LocPL* is not statistically significantly better than *PL*, with p-values of 0.1895 and 0.3426, respectively (Fig.6). However, if we restrict our analysis for the $\alpha 6\beta 4$ Integrin pathway to the first 50 windows we note that the *PosFrac* values for *LocPL* is higher than for *PL*. The first 50 windows corresponds to the first 5,000 paths in the $\alpha 6\beta 4$ Integrin pathway reconstruction. Strikingly, *LocPL* captures nearly as many positive interactions by path number 5,000 as *PL* captures by path 20,000, denoted by the vertical dashed lines in the PR curves in Fig. 4. On the other hand, the recall of *PL* after 5,000 paths is about half its recall after 20,000 paths. *LocPL* discovers the majority of positives much earlier than *PL*; thus, while the *PosFrac* values are low after path number 5,000 for *LocPL*, there are not many new interactions to discover past this point.

c) Assessment of Aggregate Pathways: To assess overall effect of *LocPL* on signaling pathway reconstructions, we considered precision and recall aggregated over the eight NetPath signaling pathways for $PLNet_2$ (Fig. 7, Left). *LocPL* shows better performance over *PL* at almost all the k values used to compute precision and recall. This improvement is striking at early values of recall, with gains of 30%, 24%, and 25% in precision at recall of 0.05, 0.1, and 0.2 respectively. While the improvement of *LocPL* is marginally significant, the aggregate F_1 score values are higher at earlier intervals for *LocPL* (Fig. 7, Right).

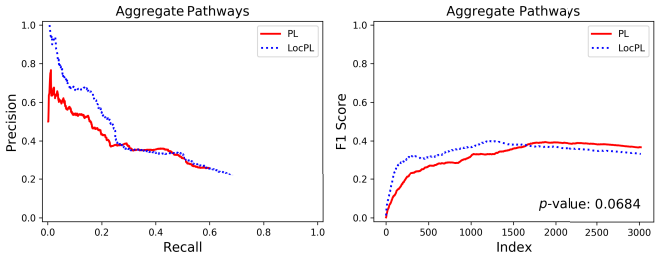

Fig. 7. Aggregate PR curve (**Left**) and F_1 score curve (**Right**) over eight signaling pathways from the NetPath database compared for *PL* and *LocPL*. The p-value is for the Mann-Whitney U test (alternative: *LocPL* $>$ *PL*).

C. Comparison of Pathway Reconstructions

LocPL provides a compartment-aware ranking of paths connecting receptors to TRs. In addition to the global and local assessments provided above, we examined the 100 top-ranking paths in the *PL* and *LocPL* pathway reconstructions for the $\alpha 6\beta 4$ Integrin, IL-7, EGFR1, and Wnt pathways. We first counted the number of paths with at least one positive interaction and the number of paths whose all interactions are positives for *PL* and *LocPL* within the first 10 and 100 paths. In all cases for all pathways, *LocPL* identifies more positive-enriched paths than *PL*.

The reconstructions, defined as the subgraph comprised of the first 100 paths, contain different numbers of nodes and edges depending on the amount of protein and interaction reuse among the paths. For the $\alpha 6\beta 4$ Integrin pathway, *LocPL* produces a smaller reconstruction compared to *PL* (in terms of nodes and edges); the opposite effect was observed for the Wnt pathway; and the IL-7 and EGFR1 pathway reconstructions are about the same size (Table II). Strikingly, a large proportion of nodes and edges in each pathway reconstruction are unique to the method (either *PL* or *LocPL*); for example, while IL-7 and EGFR1 pathway reconstructions are similar in size for each method, each reconstruction contains between 45–52% of unique nodes and between 65–78% unique edges.

We also generated networks for each pathway reconstruction (IL-7 is shown Fig. 8). The signaling flow constraints on *LocPL* paths imply two features about these networks: i) the node colors should change from red (membrane) to green (cytosol) to blue (nucleus), and ii) no paths of length one are allowed.

The Interleukin-7 (IL-7) pathway plays a major role in

TABLE II
NUMBER OF NODES AND EDGES FOR THE FIRST 100 PATHS IN EACH PATHWAY RECONSTRUCTION. "% UNIQUE" APPEAR IN ONE METHOD.

Pathway	Method	# Nodes (% Unique)	# Edges (% Unique)
$\alpha 6\beta 4$ Integrin	*PL*	89 (47%)	182 (52%)
	LocPL	56 (10%)	141 (38%)
EGFR1	*PL*	52 (50%)	122 (78%)
	LocPL	53 (52%)	112 (76%)
IL7	*PL*	40 (45%)	117 (65%)
	LocPL	41 (48%)	116 (65%)
Wnt	*PL*	43 (44%)	90 (72%)
	LocPL	65 (95%)	134 (81%)

the development and proliferation of the T cells [19]. The *PL* reconstruction of IL-7 contains many proteins that are predicted to be at the membrane, including other interleukin receptors (IL2RA and IL2RB), Thyroid Stimulating Hormone Receptor (TSHR), Leptin Receptor (LEPR), Fibroblast Growth Factor Receptor 1 (FGFR1), and a signal transducer (IL6ST). In addition to ESR1, Prolactin Receptor (PRLR) appears in the *PL* reconstruction as another nuclear receptor. Many of these proteins are "downstream" of Janus Kinase 2 (JAK2), a non-receptor tyrosine kinase. However, since JAK2 is predicted to appear slightly more in the cytosol than the other compartments, paths that contain receptors after JAK2 have a lower signaling score in the *LocPL* reconstruction. In Fig. 8(bottom), the JAK family tend to appear as the third or fourth protein in the path, with all interleukin receptors "upstream" of JAK1, JAK2, and JAK3. *LocPL*'s pathway reconstruction focuses more on interleukin signaling and effects, compared to *PL*'s reconstruction which includes potential hypotheses about interleukin's influence on other signaling pathways.

D. Pathway Reconstructions in the HIPPIE Interactome

We also applied *LocPL* to HIPPIE (Human Integrated Protein Protein Interaction rEference), a repository of 16,707 proteins and weighted 315,484 PPIs (version 2.1, July 18^{th}, 2017). We do not show these results due to space constraints, but summarize what we found. We examined four NetPath signaling pathways (EGFR1, IL-7, IL3, and IL6) that contained enough NetPath positive interactions within HIPPIE to assess *LocPL* [16] . Earlier paths of *LocPL* had more positive interactions than those of *PL*, though this trend was not statistically significant according to the MWU test. Taking the IL6 pathway as an example, the first 500 paths by *PL* contained only 4 paths with at least one positive interaction, leading to a recall of 0.22. On the other hand, the first 500 paths for *LocPL* contained 58 paths with at least one positive interaction, achieving a recall value of 0.67. In addition, final recall value is higher for *LocPL* meaning higher proportions of positives are discovered.

IV. Discussion

We present *LocPL*, an automatic signaling reconstruction algorithm that incorporates information about protein localization within the cell. Previous reconstructions contained many tied paths. *LocPL* overcomes this obstacle with a computational framework that favors paths that follow specific assumptions of signaling flow. This framework includes filtering interactions based on their predicted interaction score and applying a dynamic program to each path that finds the most likely series of cellular compartments that are consistent with the model of signaling flow.

Using a new interactome, $PLNet_2$, we have shown that *LocPL* pathway reconstructions for four pathways are more enriched with positive interactions than paths computed by *PL*. Precision of *LocPL* dominates the precision of *PL* at nearly every value of recall (Fig. 4), and the resulting F_1 scores are significantly better for *LocPL* in three of the four

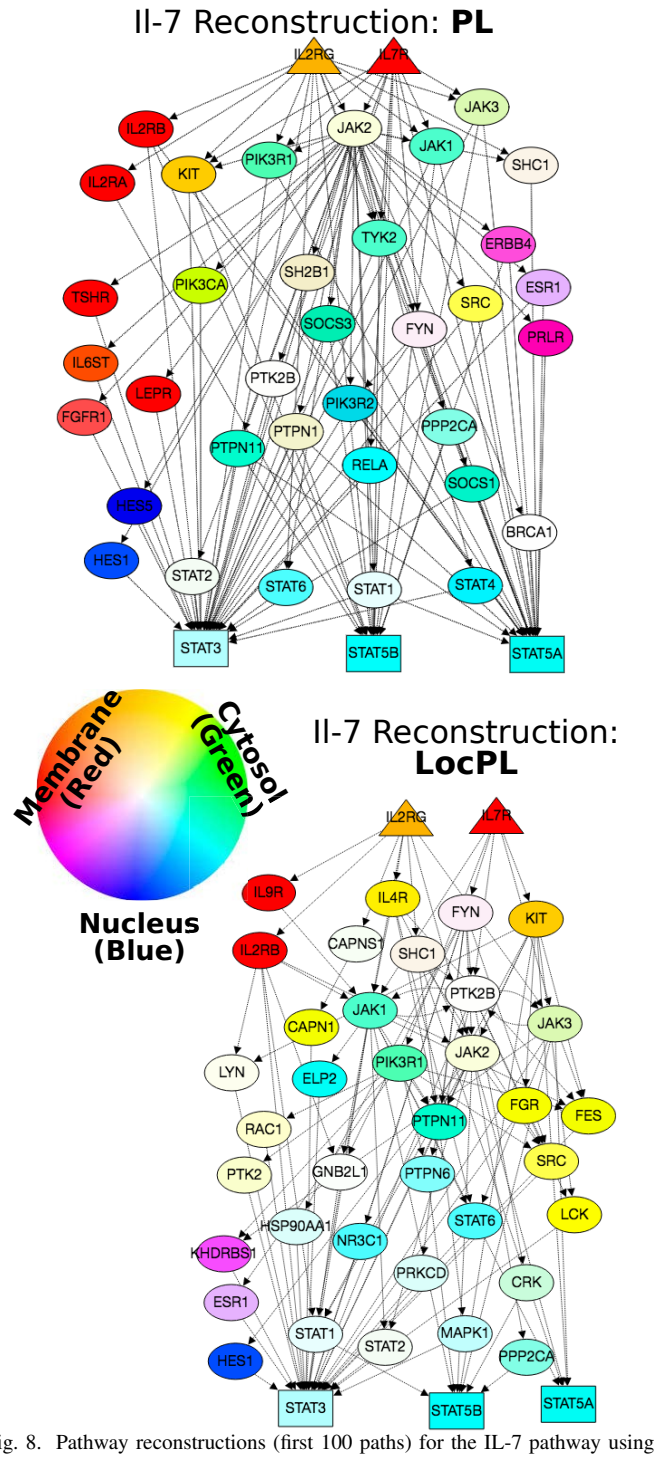

Fig. 8. Pathway reconstructions (first 100 paths) for the IL-7 pathway using *PL* (top) compared to *LocPL* (bottom). Receptors are labeled as triangles; TRs are rectangles, intermediary proteins are ellipses. Color denotes compartment localization; proteins may belong to multiple compartments (and will be lighter shades). Networks generated using GraphSpace [20].

pathways (Fig. 5). *LocPL* dramatically improves precision at low values of recall across eight signaling pathways, though this difference is nearly significant by the MWU test (Fig. 7). Moreover, *LocPL* ensures that the constituting interactions, from a receptor to a TR, are spatially-coherent within the different cellular compartments. This feature increases the number of paths that contain positives early in the pathway

reconstruction, which supports our hypothesis that *LocPL* locally promotes paths with higher proportions of positives up in the k-shortest paths list.

In this work, we chose to impose an ordering on a subset of the available compartments from ComPPI (*ExtMem*, *Cytosol*, and *Nucleus*). There are many ways to impose a compartmental ordering of signaling flow to capture other features of signaling, including mitochondria-dependent signaling, nuclear receptor signaling and extracellular signaling. *LocPL* is generalizable to different signaling models, as long as the user specifies compartment relationships in a memoryless manner. To illustrate this point, we developed a model of signaling that also includes the mitochondria compartment. To highlight the versatile behavior of interaction the mitochondria have with the other subcellular components, we allowed the mitochondria to have direct interactions with, and to be of order-variant regarding, the other three cellular compartments in our signaling model. There were very few changes in the results when we included the mitochondria into our signaling model, most likely due to the relatively few number of proteins in $PLNet_2$ that had non-zero *Mitochondria* localization scores (data not shown).

In addition to the precision and recall assessment used previously by PathLinker, we proposed a measure to assess individual paths in terms of proportion of positive signaling interactions. This measure offers complementary insights to the pathway reconstructions by *PL* and *LocPL*. PR curves demonstrate how quickly positive interactions are recovered in a reconstruction, but do not consider the fact that many paths may contain the same positives. The path-based measure considers the proportion of positives within a set of paths, demonstrating that some sets of paths are enriched for positive interactions that may have appeared in a higher-ranked path. This signal cannot be captured with the previous global measure.

The framework that we have developed may be extended to other graph-theoretic approaches that return subnetworks of directed structure with an associated reconstruction score, such as trees [9], [21], [22]. Our approach encourages the enumeration of many tied results, since incorporating protein compartment information will help break these ties with biologically relevant information. In addition, we anticipate to develop the technique to compare paths in the context of different conditions, such as dysregulated signaling as a result of disease or tissue-specific signaling.

REFERENCES

[1] M. Schaefer, J. Fontaine, A. Vinayagam, P. Porras, E. Wanker, and M. Andrade-Navarro, "HIPPIE: Integrating protein interaction networks with experiment based quality scores.," *PLoS ONE*, vol. 7, no. 2, p. e31826, 2012.

[2] G. Alanis-Lobato, M. Andrade-Navarro, and M. Schaefer, "HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks.," *Nucleic Acids Research*, vol. 45, pp. D408–D414, 2017.

[3] B. Aranda, H. Blankenburg, S. Kerrien, F. Brinkman, A. Ceol, *et al.*, "PSICQUIC and PSISCORE: accessing and scoring molecular interactions," *Nat. Methods*, vol. 8, pp. 528–529, Jun 2011.

[4] P. Hornbeck, J. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, *et al.*, "PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse," *Nucleic Acids Res.*, vol. 40, pp. D261–270, Jan 2012.

[5] M. Steffen, A. Petti, J. Aach, P. D'haeseleer, and G. Church, "Automated modelling of signal transduction networks," *BMC bioinformatics*, vol. 3, no. 1, p. 34, 2002.

[6] J. Scott, T. Ideker, R. M. Karp, and R. Sharan, "Efficient algorithms for detecting signaling pathways in protein interaction networks," *Journal of Computational Biology*, vol. 13, no. 2, pp. 133–144, 2006.

[7] A. Ritz, C. Poirel, A. Tegge, N. Sharp, K. Simmons, A. Powell, *et al.*, "Pathways on demand: automated reconstruction of human signaling networks," *npj Systems Biology and Applications*, vol. 2, p. 16002, 2016.

[8] J. Supper, L. Spangenberg, H. Planatscher, A. Drager, A. Schroder, and A. Zell, "BowTieBuilder: modeling signal transduction pathways," *BMC Syst Biol*, vol. 3, p. 67, Jun 2009.

[9] N. Tuncbag, A. Braunstein, A. Pagnani, S. Huang, J. Chayes, *et al.*, "Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem," *J. Comput. Biol.*, vol. 20, pp. 124–136, Feb 2013.

[10] Y. Sun, C. Ma, and S. Halgamuge, "The node-weighted Steiner tree approach to identify elements of cancer-related signaling pathways," *BMC Bioinformatics*, vol. 18, p. 551, Dec 2017.

[11] E. Yeger-Lotem, L. Riva, L. Su, A. Gitler, A. Cashikar, *et al.*, "Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity," *Nat. Genet.*, vol. 41, pp. 316–323, Mar 2009.

[12] I. Nassiri, A. Masoudi-Nejad, M. Jalili, and A. Moeini, "Discovering dominant pathways and signal-response relationships in signaling networks through nonparametric approaches," *Genomics*, vol. 102, no. 4, pp. 195 – 201, 2013.

[13] J. Sun, M. Zhao, P. Jia, L. Wang, Y. Wu, *et al.*, "Deciphering signaling pathway networks to understand the molecular mechanisms of metformin action," *PLOS Computational Biology*, vol. 11, pp. 1–35, 06 2015.

[14] E. Paull, D. Carlin, M. Niepel, P. Sorger, D. Haussler, and J. Stuart, "Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (TieDIE)," *Bioinformatics*, vol. 29, no. 21, pp. 2757–2764, 2013.

[15] D. Veres, M. Gyurko, B. Thaler, K. Szalay, D. Fazekas, *et al.*, "ComPPI: a cellular compartment-specific database for protein–protein interaction network analysis," *Nucleic Acids Research*, vol. 43, no. D1, pp. D485–D493, 2015.

[16] K. Kandasamy, S. Mohan, R. Raju, S. Keerthikumar, G. Kumar, *et al.*, "NetPath: a public resource of curated signal transduction pathways," *Genome Biology*, vol. 11, p. R3, Jan 2010.

[17] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Res.*, vol. 45, pp. D353–D361, Jan 2017.

[18] A. Paz, Z. Brownstein, Y. Ber, S. Bialik, E. David, *et al.*, "SPIKE: a database of highly curated human signaling pathways," *Nucleic Acids Research*, vol. 39, pp. D793–D799, Jan. 2011.

[19] A. Plumb, A. Sheikh, D. Carlow, D. Patton, H. Ziltener, and N. Abraham, "Interleukin-7 in the transition of bone marrow progenitors to the thymus," *Immunology And Cell Biology*, vol. 95, pp. 916–924, Aug 2017.

[20] A. Bharadwaj, D. Singh, A. Ritz, A. Tegge, C. Poirel, *et al.*, "Graphspace: stimulating interdisciplinary collaborations in network biology," *Bioinformatics*, vol. 33, no. 19, pp. 3134–3136, 2017.

[21] N. Yosef, L. Ungar, E. Zalckvar, A. Kimchi, M. Kupiec, *et al.*, "Toward accurate reconstruction of functional protein networks," *Molecular Systems Biology*, vol. 5, no. 1, 2009.

[22] M. Bailly-Bechet, C. Borgs, A. Braunstein, J. Chayes, A. Dagkessamanskaia, *et al.*, "Finding undetected protein associations in cell signaling by belief propagation," *Proceedings of the National Academy of Sciences*, vol. 108, no. 2, pp. 882–887, 2011.