Synthesizing Signaling Pathways from Temporal Phosphoproteomic Data

Ali Sinan Köksal*, Anthony Gitter*, Kirsten Beck, Aaron McKenna, Saurabh Srivastava, Nir Piterman, Rastislav Bodík, Alejandro Wolf-Yadlin, Ernest Fraenkel, Jasmin Fisher

Advances in proteomic measurements reveal that even the best-curated pathways fail to capture a large fraction of signaling events. Here we propose a synthesis approach to produce precise models of signal transduction from temporal phosphoproteomic data. We first integrate the time series data with a protein-protein interaction network to produce an initial undirected graph. Using program synthesis techniques, we exhaustively explore all possible signaling pathways that are consistent with the proteomic data and the initial graph without explicitly enumerating all models. These pathways must satisfy several logical constraints. Most notably, a chain of events initiating at the source of the stimulation, the epidermal growth factor (EGF) in this case, must explain the activation or inhibition of each phosphorylated protein. In addition, the timing of all events must agree with the temporal data such that upstream proteins are not activated or inhibited after their downstream neighbors. This approach identifies parts of the network that are consistent with all possible pathway models. We are able to determine the direction of interactions, whether edges activate or inhibit, and the times at which proteins are activated.

Using new mass spectrometry data of the temporal EGF response in EGFR Flp-In HEK-293 cells, we show that nearly all proteins that change significantly in phosphorylation (89 to 98% depending on the database) are absent from canonical maps of the epidermal growth factor receptor (EGFR) signaling. Our computational approach reconstructs and summarizes all valid pathway models that explain how proteins are activated or inhibited by EGF. Collectively, these models account for 83% of the significant proteins and contain 413 protein-protein interactions, of which 200 can be confidently assigned a direction. In all cases where we predict a directed interaction between two EGFR pathway nodes, the prediction is correct. We use three natural language processing (NLP) tools to search for literature support for 54 predicted pathway edges that are peripheral to known EGFR pathway interactions. Manually verifying the results, we find that the direction is correct for 15 of the 16 predictions for which there is a definitive direction in the literature. Overall, of the 200 predicted directed pathway edges, 82 are supported by the canonical EGFR pathway, NLP, or kinase-substrate interactions (whose directions are included as prior knowledge). We are presently testing several predictions experimentally by assessing whether kinase inhibitors disrupt phosphorylation of the predicted substrate at the specific times proposed by our model. In summary, our computational approach identifies many previously unrecognized components of a well-studied signaling pathway. Our technique is broadly applicable to systems where dynamic proteomic data is

available and has great potential for constructing pathway maps in conditions that alter classic signaling cascades, such as in diseased cells.

*equal contribution

Keywords: protein-protein interaction network, program synthesis, Steiner forest, time series data, mass spectrometry

Category: Single-cell biology, proteomics, signaling