Predictive Analytics of Phosphoproteins in Breast Cancer Cells

Stephen Obol Opiyo
Molecular and Cellular Imaging Center – Columbus , Ohio Agricultural Research and Development Center

Abstract

Cells respond to their environment by activating signaling networks that trigger processes such as growth, survival, apoptosis (cell death), and migration. Post-translational modifications, notably phosphorylation, play a key role in signaling. In cancer cells, signaling networks frequently become compromised, leading to abnormal behaviors and responses to external stimuli. Advancing our understanding of how these networks are deregulated across cancer cells will ultimately lead to more effective treatment strategies for patients. There is an urgent need for computational approaches that can characterize causal signaling networks using data acquired in a specific background or context of interest. There is also a need to address the related task of predicting dynamical trajectories in specific contexts and under specific perturbations. In the Dialogue on Reverse Engineering Assessment and Methods (DREAM) challenge 8, participants were given breast cancer data and asked to use the data to build models that can predict trajectories of protein levels following inhibitor perturbation(s) not included in the training data. Data were acquired under 8 stimuli (Serum, PBS, EGF, Insulin, FGF1, HGF, NRG1, IGF1), and inhibition of network nodes by one of 3 inhibitors (AKT, [AKT + MEK], [FGFR1, FGFR3]) or DMSO vehicle control (cells were serum-starved and pre-treated with inhibitor prior to ligand stimulation). The experiment was carried out on 4 breast cancer cell lines (MCF7, UACC812, BT20, and BT549), with abundance of ~45 phosphoproteins measured at 7 time points post-stimulus. This poster presents the results of the predictions using multivariate statistic (partial least squares regression), compared to other methods.

Objective

The objective is to build dynamical models that can predict trajectories of phosphoproteins. An important emphasis is on the ability of models to generalize beyond the training data by predicting trajectories under perturbations not included in the training data.

Materials and Methods

Experimental breast cancer proteomic data

Dataset was generated using Reverse Phase Protein Array (RPPA) quantitative proteomics technology. The experiment design is shown in Figure 1.

- Data were acquired under 8 stimuli (Serum, PBS, EGF, Insulin, FGF1, HGF, NRG1, IGF1)
- Inhibition by one of 3 inhibitors (AKT, [AKT + MEK], [FGFR1, FGFR3])
- 4 breast cancer cell lines (MCF7, UACC812, BT20, BT549)
- Abundance of 45 phosphoproteins (proteins phosphorylated at specific sites) measured.
- Measured at 7 time points post-stimulus (0 min, 5 min, 15 min, 1 hr, 2 hr, 4 hr).

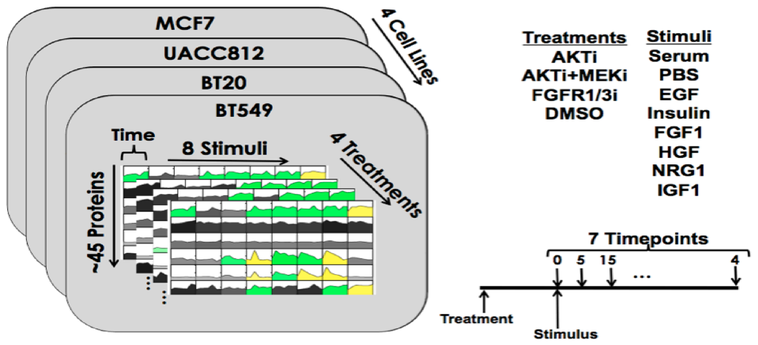

Figure 1. An overview of the experimental design to generate the dataset.

Method

Partial least squares regression (PLS) Figure 2

- PLS takes into account latent structure in both datasets
- X_i are the k explanatory variables
- Y_j are the p dependent variables
- The model is linear $y_{nj} = \sum_{i=0}^{k} \beta_i x_{ni} + \varepsilon_{nj}$
- β_i are found differently from linear regression

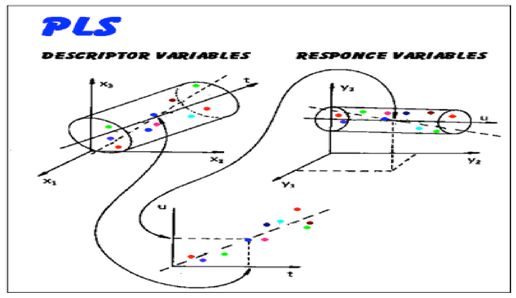

Figure 2. The overview of PLS regression

Results

Participants	Affiliation	Method	Median z-score	p-value (median z-score)
GuanLab	University of Michigan	Singular Value Decomposition based on Markov	-3.5785	1.73E-04
Stochastic Chaos	New Mexico State University	Time series Analysis	-3.59	1.65E-04
Opiyo	**Ohio State University**	**Partial least square regression**	**-3.5989**	**1.60E-04**
JCheng	Computational Biology, Zurich, Switzerland	Support vectot machine	-2.933	1.68E-03
sakev	Columbia University	Gradient Boosted Decision Trees	-3.0633	1.09E-03
Dynamo Bios	Northwestern University	Bayesian Inference Method	-1.9511	0.0255
Hatric	No information	No information	-1.622	0.0524
CGR	Chinese Academy of Medical Sciences	Time series Analysis	-1.1357	0.128
HD_Systems	Ruprecht-Karls-University of Heidelberg, Germany	Correlation-based Elimination	2.1213	0.983053792
StuartLab	No information	Time series Analysis	1.6465	0.950164899
No information	No information	No information	48.0036	1
Freya	No information	No information	142.6412	1
SBIT	University of Padova, Italy	Sparse Ordinary Differential Equations	185.3177	1

Conclusions

- PLS method effectively predicted phosphoproteins in cancer cells that were not included in the training data.
- PLS method was among the best methods for phosphoproteins predictions.

References

- Neve et al. 2006, Cancer Cell.
- Hill et al, (The HPN-DREAM Consortium [Opiyo, SO]) (2016). Inferring causal molecular networks: empirical assessment through a community-based effort. Nature Methods. doi:10.1038/nmeth.3773
- Federica Eduati et al. (The NIEHS-NCATS-UNC DREAM Toxicogenetics Collaboration [Opiyo, S.O]) (2015). Prediction of human population responses to toxic compounds by a collaborative competition. Nature Biotechnology 33, 933–940