Proteomic Screening and Lasso Regression Reveal Differential Signaling in Insulin and Insulin-like Growth Factor I (IGF1) Pathways*

Cemal Erdem‡§, Alison M. Nagle¶¶, Angelo J. Casa**¶¶, Beate C. Litzenburger*, Yu-fen Wang**, D. Lansing Taylor‡§, Adrian V. Lee¶¶, and Timothy R. Lezon‡§§

Insulin and insulin-like growth factor I (IGF1) influence cancer risk and progression through poorly understood mechanisms. To better understand the roles of insulin and IGF1 signaling in breast cancer, we combined proteomic screening with computational network inference to uncover differences in IGF1 and insulin induced signaling. Using reverse phase protein array, we measured the levels of 134 proteins in 21 breast cancer cell lines stimulated with IGF1 or insulin for up to 48 h. We then constructed directed protein expression networks using three separate methods: (i) lasso regression, (ii) conventional matrix inversion, and (iii) entropy maximization. These networks, named here as the time translation models, were analyzed and the inferred interactions were ranked by differential magnitude to identify pathway differences. The two top candidates, chosen for experimental validation, were shown to regulate IGF1/insulin induced phosphorylation events. First, acetyl-CoA carboxylase (ACC) knock-down was shown to increase the level of mitogen-activated protein kinase (MAPK) phosphorylation. Second, stable knock-down of E-Cadherin increased the phospho-Akt protein levels. Both of the knock-down perturbations incarcerated phosphorylation responses stronger in IGF1 stimulated cells compared with insulin. Overall, the time-translation modeling coupled to wet-lab experiments has proven to be powerful in inferring differential interactions downstream of IGF1 and insulin signaling, in vitro. Molecular & Cellular Proteomics 15: 10.1074/mcp.M115.057729, 3045–3057, 2016.

This is an open access article under the CC BY license.

From the ‡Department of Computational & Systems Biology, University of Pittsburgh, Pittsburgh, Pennsylvania; §University of Pittsburgh Drug Discovery Institute, University of Pittsburgh, Pittsburgh, Pennsylvania; ¶Department of Pharmacology & Chemical Biology, University of Pittsburgh, Pittsburgh, Pennsylvania; ¶¶Women’s Cancer Research Center, University of Pittsburgh Cancer Institute, Pittsburgh, Pennsylvania; **Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, Texas; ††Department of Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania

Received December 28, 2015, and in revised form, June 23, 2016
Published, MCP Papers in Press, June 30, 2016, DOI 10.1074/mcp.M115.057729

Author contributions: DLT, ALV and TRL designed the research; AMN, AJC, BCL and YW performed experiments; CE constructed all computational models; CE, AMN, DLT, AVL and TRL analyzed data and wrote the paper.

Insulin (Ins) and type I insulin-like growth factor (IGF1) receptors (InsR and IGF1R, respectively) are receptor tyrosine kinases that are expressed in almost all types of cells. Signaling through InsR and IGF1R initiates a phosphorylation cascade that drives cell growth and proliferation (1–8). Overexpression of these receptors is correlated with higher breast cancer risk (9–15) and has been shown to influence tumorigenesis, metastasis, and resistance to existing forms of cancer therapy (10, 16, 17). IGF1 receptor blockade can slow tumor growth and metastasis, but the receptor has proven to be difficult to target specifically (18–20). A confounding factor in developing therapies targeting these receptors is their high sequence (~60%) and structural homology. IGF1R and InsR are able to form functional hybrids, and each can partially compensate for the loss or suppression of the other (1–4, 21–24). Moreover, it has been shown that IGF1R signaling is one mechanism of resistance to conventional hormonal therapy (19, 25–29). Understanding the relationships between IGF1R and InsR signaling cascades and their combinatorial role in cancer is crucial to developing better diagnostics and personalized treatments for cancer.

© 2016 by The American Society for Biochemistry and Molecular Biology, Inc. This paper is available on line at http://www.mcponline.org

The abbreviations used are: Ins, Insulin; ACC, Acetyl-coenzyme-A carboxylase; Akt, Rac-alpha serine/threonine-protein kinase (Protein kinase B); AMP: Adenosine monophosphate; AMPK, 5′-AMP-activated protein kinase catalytic subunit alpha-1; ATM, Serine-protein kinase Ataxia telangiectasia mutated; ATP, Adenosine triphosphate; COX2, Prostaglandin G/H synthase 2 (Cyclooxygenase-2); E-Cadherin, Epithelial cadherin; EGFR, Epidermal growth factor receptor; EMT, Epithelial-mesenchymal transition; ENM, Elastic network model; HER2, Receptor tyrosine-protein kinase erbB-2; IGF1, Insulin-like growth factor I; IGF1R, IGF1 receptor; InsR: Insulin receptor; LOOCV, Leave-one-out cross-validation; MAPK, Mitogen-activated protein kinase; MaxEnt, Entropy maximization; mTORC1, Mammalian target of rapamycin complex 1; NADPH, Nicotinamide adenine dinucleotide phosphate (reduced form); NF-kB, Nuclear factor kappa B; ns: non-significant; p70S6K, Ribosomal protein S6 kinase beta-1; PI3K, Phosphatidylinositol 4,5-bisphosphate 3-kinase; PKCa, Protein kinase C alpha type; PTEN, Phosphatidylinositol 3,4,5-trisphosphate 3-kinase; R, Pearson correlation coefficient; RAS, GTPase Ras; RB1, Retinoblastoma-associated protein; RPPA, Reverse phase protein array; S6: Ribosomal protein S6; Scr, Scrambled; S.E., Standard error of the mean; SFM, Serum free medium; std: Standard deviation; SVD, Singular value decomposition; TSC2, Tuberin (Tuberous sclerosis 2 protein); TT, Time translation; Vhc, Vehicle; WT, Wild-type.
In order to gain insight into cellular signaling networks, a method of tracking the evolution of multiple cellular features in time is required. Advances in experimental procedures and instruments have led to the development of proteomics approaches such as mass spectrometry (8) and live cell imaging (30) that enable the quantification of the levels of multiple proteins across large numbers of samples. Similarly, reverse phase protein array (RPPA) is a high-throughput technique that provides quantification of expression levels of total proteins as well as specific phosphorylated forms (31). Researchers have already used the power of RPPA toward biomarker prediction (32), classification of responses to diverse inputs (33), and study of drug responsiveness (34). In this study, RPPA data obtained from IGF1 and Ins stimulated breast cancer cell lines (n = 21) over a 48 h time course was utilized to construct models of time translation to characterize important differential dynamics and interactions present downstream of two structurally similar yet functionally different growth factors.

High throughput data sets have been the subject of computational modeling for well over a decade. Early linear models applied to gene expression data were able to extract clear signals from superficially noisy time-series data. Conventional tools and methods for studying temporal dynamics of genomic data (35) and pathways (36) often rely on sophisticated machine learning techniques that provide high accuracy at the cost of clarity. The utility of simpler linear models is that their results can be interpreted in terms of known biology. Whereas complex models may tend to obfuscate their inner workings, linear models connect components directly to each other in an unambiguous fashion. Application of linear elastic network models (ENM) in structural biology, for instance, enables one to analyze very large systems by lowering complexity (37–39). Coarse-grained motions and equilibrium dynamics can be recapitulated with very high accuracy using reductionist models. These and other studies (40–42) have shown that simple, cheaper, and faster approaches are valuable in beginning to understand and apply current knowledge to drive better models of biological systems.

Over the last few decades, the progression from static to time-series data has enabled researchers to infer rates of change of state variables, to find causal interactions in-between, and to study collective system dynamics. The algorithms and protocols to analyze the time-dependent data generated have been expanding in parallel, especially with the invention and development of microarray technology, and a variety of tools and methods are available for studying temporal dynamics (35). Although most of the techniques are based on the genomic data, they are applicable to other types of temporal data as well. These methods span a large set of approaches, including both linear and nonlinear models (43–50). Linear models, like ones performed in studies focusing on construction of circadian gene regulation or interaction network topologies (48, 51), serve as tools of easy implementation and interpretation.

Here, we use a simple yet intuitive linear model to explore the time evolution of protein expression profiles in breast cancer cell lines stimulated with IGF1 or Ins. Using the lasso algorithm trained on the time-series RPPA expression profiles, we construct temporal networks of effective interactions between signaling proteins. Analyzing the resultant networks, we find novel temporal differences in cellular response to IGF1 and insulin stimulation. The approach generates a set of predictions that are validated using wet-lab perturbation experiments.

**MATERIALS AND METHODS**

**Cell Culture and Reverse Phase Protein Array—**The initial data set (supplemental Table S1, also deposited in NCBI’s Gene Expression Omnibus (GEO) (52) database with series accession number GSE80233) used here was generated at the RPPA Core Facility (MD Anderson Cancer Center, Houston, TX) (34, 53, 54). Twenty-one breast cancer cell lines (supplemental Table S2) were cultured in DMEM, DMEM/F-12, RPMI, L15 or McCoy’s 5A media supplemented with 10% FBS or horse serum according to the conditions provided by Neve et al. (55). Cells were plated in triplicate 6 cm dishes per treatment time point and then switched to serum-free medium (SFM). Cells were serum starved for 24 h, then treated with 10 nM IGF1 (Novozymes, Franklinton, NC) or Ins (Sigma, St. Louis, MO) for 5, 10m, 30m, 6h, 24h, and 48h. Cells treated with vehicle (Vhc), the serum-free condition were collected at 5m, 24h, and 48h. Cell lysates were harvested by washing in ice-cold PBS twice, and lysing immediately in RPPA lysis buffer with protease inhibitor (1% Triton X-100, 50 mM Hepes, 150 mM NaCl, 1.5 mM MgCl2, 1 mM EGTA, 100 mM NaF, 10 mM NaPi, 10% Glycerol, 1 mM Na3VO4, 1 mM PMSF, 10 μg/ml Aprotinin). Lysates were sonicated and centrifuged at 14,000 rpm at 4 °C for 10 min. Protein concentrations were determined by BCA. Then, lysates were mixed with 4X SDS/2-MA sample buffer to have final concentration of 1 μg/ml. The lysates were boiled for 5 min and diluted twofold five times. The diluted lysates were arrayed on nitrocellulose-coated FAST slides (GE Healthcare, Chicago, IL). Each slide was probed with a primary antibody and a biotin-conjugated secondary antibody. Levels of 134 proteins or phospho-proteins were measured using corresponding antibodies. The signal, amplified by Dako-Cytomation-catalyzed system (Dako, Carpinteria, CA), was visualized by DAB colorimetric reaction. The slides were scanned and quantified using Microvion (VigenTech, Carlisle, MA) software. Supercurve fitting was used for loading correction of each sample (56).

**Quality Control of RPPA Antibodies—**We performed an internal quality control on the specificity and validity of the 134 antibodies used in the RPPA data generation. This step included a condition of obtaining a single or dominant band for the antibody in question on Western blots ran for the dynamic range conditions of RPPA data. Then, a Pearson correlation coefficient of 0.8 between RPPA and WB results was used as cutoff to determine if the antibody is specific enough for the intended target. The RPPA Core Facility (57) also considered intra- and inter-slide reproducibility for the antibodies. They update their list of available reliable antibodies regularly based on experiences in hundreds of conditions and cell lines. Finally, 26 of the antibodies were labeled not specific enough under the specific experimental conditions, and are excluded from all subsequent analyses.

**siRNA Knock-down and Immunoblotting—**The HCC1569 and MCF7 cell lines were acquired from ATCC and grown in RPMI+10% FBS. ACACA and ACACB (gene names for acetyl-coenzyme-A car-
boxylase isoforms) pooled siRNAs (Dharmacon, Lafayette, CO) were reverse transcribed with Lipofectamine RNAi MAX, following the manufacturer’s protocol (Life Technologies, Waltham, MA), with a final concentration of 25 nM. T47D cells acquired from ATCC were stably transfected with shRenilla and shCDH1 (CDH1: E-Cadherin gene name) constructs using a retroviral infection and were grown in RPMI+10% FBS with 1 μg/ml puromycin. Cells were serum starved for 24 h then stimulated with 10 nM IGF1 or insulin for 5 min, 10 min, 30 min, and 6 h time points. 10 nM HCl was used as a vehicle control. Protein lysates were collected with RIPA buffer containing protease and phosphatase inhibitors, and standard immunoblot techniques were used to evaluate protein expression. The pMAPK (Thr202/Tyr204, Cell Signaling, Danvers, MA; 1:1000), MAPK (Cell signaling, 1:1000), pAkt (Ser473; Cell signaling; 1:1000), Akt (Cell signaling; 1:1000), E-cadherin (BD Biosciences, San Jose, CA; 1:1000), ACC (Cell Signaling; 1:500) and β-actin (Sigma; 1:5000) antibodies were used. RNA was isolated using the Illustra RNeasy Mini Kit (GE Healthcare) and reverse transcription was done following the iScript cDNA synthesis kit protocol (BioRad, Hercules, CA). The SYBR green supermix protocol (BioRad) was used for qRT-PCR. Primers for ACACA (5' GAATCGGACTGGCAGAAG (fwd) and 5' GCAGAGTGCCGTCCTGTTCCAG (rev)) were used to evaluate siRNA transfection efficiency.

**Construction of Linear Statistical Models**

**Data Preprocessing**—To reduce noise, we selected only proteins with large expression-level variance. We calculated the variance of each (phospho)-protein expression across all conditions and identified a variance cutoff that separates our antibody set into 43 high-variance proteins and 65 low-variance proteins. The 0.42 cutoff is similar to each other than the mean of triplicates across different conditions.

**Supplementary Fig. S1**, so that small changes in the cutoff value do not alter the number of proteins retained. We also showed that the data from triplicates are more similar to each other than the mean of triplicates across different conditions (supplemental Fig. S2). The Kolmogorov-Smirnov distance between the standard deviations of the triplicates and the standard deviation of the mean of triplicates across conditions is 0.98 (supplementary Fig. S2). In addition, when we performed one-way ANOVA on the filtered RPPA data for every condition separately, the results indicated that the mean of triplicate measurements across proteins are significantly different (highest p value ~ 6e-26). As a result, we only used the mean of triplicate experiments in all subsequent analyses. The individual time point data sets (each of size (43 proteins) × (21 cell lines)) were then column centered, row-centered, and row-normalized in a step that causes minimal information loss.

The time-translation matrices constructed here satisfy the equation $X(t_i) = \mathbf{T} X(t_0)$, where the expression vector $X(t)$ contains the levels of $P$ proteins measured at time point $t$. The time translation matrix $\mathbf{T}$ is a square matrix that transforms protein expression levels from an early time point $t_0$ to a later one $t_i$. As any matrix that exactly transforms all the data from one time point to another is considered a valid time translation matrix, the problem of defining $\mathbf{T}$ is underdetermined if the number of measured proteins exceeds the number of samples. We investigated three different methods of calculating $\mathbf{T}$ from the data, each representing a constrained optimization.

**Pseudoinversion**—If we compile the $N$ expression vectors at time $t_0$ into the $P \times N$ matrix $X(t_0)$ and construct $X(t_i)$ similarly, then $\mathbf{T}$ is given by

$$\mathbf{T}_{ij} = X(t_i)^{-1} X(t_0)$$

(Eq. 1)

where $X^{-1}(t_0)$ is the pseudoinverse of $X(t_0)$ when $N < P$, as in the present case. The time translation matrix calculated this way minimizes the contribution to $\mathbf{T}$ from the nullspace of $X(t_0)$. The steps of model, called SVD modeling, construction are depicted in Fig. 1A.

**Entropy Maximization**—We can alternatively assume a distribution that maximizes entropy subject to the constraints of the data. We write the equation of probability for the expression value distributions:

$$p(x(t_i), x(t_j)) = \frac{1}{Z} \exp \left\{ -\frac{1}{2} (x^T(t_i) x^T(t_j) M^{-1} x)(t_i, t_j) \right\}$$

(Eq. 2)

where $\mathbf{M}$ is the (pseudo)inverse of the joint-covariance matrix and $Z$ is the partition sum and equal to $\langle 2\pi^\top \mathbf{M}^{-1/2} \rangle$. $\mathbf{M}$ is the matrix of effective interactions between the components of the expression matrices. The steps followed for time-translation matrices construction from the $\mathbf{M}$ matrix are also given in Fig. 1A, route 1-2-5-6-7 (the MaxEnt model). The details of model construction are given in Supplementary Materials.

**Lasso Regression**—The time translation (TT) matrix $\mathbf{T}$ is obtained by an iterative application of the lasso algorithm (58, 59). For each row $k$ of the data matrix, we calculate one row of the time translation matrix $\mathbf{T}$ and one component of the offset vector $\delta$ by numerically minimizing the quantity

$$\sum_i^m |x(t_0)_i| - \sum^c_j T_{kj} x(t_0)_j - \delta_k y + \lambda \sum_i^m |T_{ij}|$$

(Eq. 3)

where the summation over $i$ minimizes the mean squared error and the last summation penalizes nonzero entries in the time-translation matrix, dictated by the positive-valued constant $\lambda$. The software used here calculates a set of $\lambda$ values at the start of the cross-validation given the inputs (60). By penalizing the L1 norm of $\mathbf{T}$, only a few nonzero coefficients are selected. The procedure is run independently for each row, and the coefficient arrays for every protein are concatenated to obtain $\mathbf{T}_L = \begin{bmatrix} T_{11} & \cdots & T_{1p} \\ \vdots & \ddots & \vdots \\ T_{m1} & \cdots & T_{mp} \end{bmatrix}$, where $\mathbf{T}_L$ represents the regression coefficient obtained for the effect of protein $j$ at time $t_i$ on the level of protein $i$ at time $t_2$. Ten thousand instances of the time translation matrix $\mathbf{T}_L$, each starting with a different random seed, are generated using the gnomet package (61). Edges that appear in at least half the matrices are considered essential and are used to select a representative matrix from the ensemble. The matrix that contains essential edges the most is selected as the representative the matrix. Route 1-8-9-10 in Fig. 1A schematically depicts the lasso model construction.

**Validation of the Lasso Models**—To validate the lasso approach and the interaction values it selects, we use randomly generated synthetic data. We find that the networks inferred from the RPPA data have fewer (typically about half as many) essential interactions than those generated from random synthetic data (supplemental Figs. S3 and S4). We further find that the distribution of interaction values inferred from synthetic data is normal (supplemental Fig. S5), whereas those inferred from the experimental data contain a few very high values and a long tail. Moreover, we show that the number of edges inferred is robust against the number networks generated, and that the number of edges in the representative network converges to the number of accepted edges when 10,000 networks are generated (supplemental Fig. S6). In addition, we find the best candidate networks always contain high magnitude interactions both in INS and IGF1 stimulation cases for full (supplemental Fig. S7) and leave-one-out cross validation (LOOCV) (supplemental Fig. S8) models. Finally, the TT matrices of lasso models for IGF1 and INS conditions have multiple interaction elements in common (supplemental Fig. S9). The lists of inferred interactions for all time translation lasso models are given in supplementary Table S3.

**Leave-one-out Cross Validation Models**—To test the performance of the three models, we further employed a leave-one-out cross validation (LOOCV) scheme (Fig. 1B). In each LOOCV step, one cell line is excluded from the training set and the three TT models are
constructed using the remaining cell lines only. The SFM expression profile of the removed cell line is used to predict its profile at the next time point. As a result, twenty-one LOOCV models are obtained for each time translation for each model. In the case of lasso models, hundred TT matrix instances are generated, and the ensemble median is selected as the representative model. Comparison of LOOCV model performances of the three methods on the same unseen data enabled us to evaluate the robustness of models.

Statistical Analysis—Standard one-way ANOVA was performed on the RPPA data, considering each condition separately. The analyses were done for the triplicate measurements of a cell line, with one stimulation condition at a specific time point. Statistical analysis on the validation experimental results was performed using two-sample t-test with significance level of $p < 0.05$. Unless otherwise indicated, the values are presented as means ± S.E. (standard error of the mean). All of the simulations and analyses were run in Matlab, versions 2014a and 2015a (MathWorks, Natick, MA).

RESULTS

Data Analysis—The data set of this study contains expression levels for 134 proteins, either in total or in a specific phosphorylated form, in 21 breast cancer cell lines stimulated in biological triplicates with IGF1 or insulin for 0–48 h (see supplemental Table S2 for information on cell line subtypes (62)). Twenty-six antibodies were excluded from this study to have only an initial set of 108 proteins (see Methods). A cursory look at the data (Fig. 2) reveals that IGF1 stimulation induces phosphorylation of IGF1R at Tyr1135 in all cell lines. Half of the cell lines also exhibit a higher phosphorylation rate of Akt and its downstream protein targets, including ribosomal protein S6. The six time-point data for each cell line show expected responses, with pAkt levels peaking around 5–10 min and decreasing gradually afterward. The cell lines MCF7, MCF10A, T47D, and ZR751 were found to be highly IGF1 responsive (see supplemental Fig. S10 for complementary Ins stimulation results).

The Lasso-based Models are Accurate and Robust—To discover signaling differences from the data, we constructed linear models of time translation (TT) using three different
methods: (1) regression coefficients selected by the lasso algorithm (58) (lasso model), (2) conventional matrix inversion (SVD model), and (3) entropy maximization (MaxEnt model). Initially, the starting RPPA data set was filtered and preprocessed. First, the raw-expression profiles were variance-ranked and the top 43 phospho or total proteins were selected for all the subsequent analyses (see supplemental Table S4). Each procedure results in a time translation matrix, \( T \), of size 43 \( \times \) 43. The element \( T_{ij} \) represents the effect that the level of protein \( j \) in unstimulated (serum-free medium, SFM) cells will have on the level of the protein \( i \) at a future time (see Methods and Fig. 1 for details of the model construction).

We obtain predicted expression profiles for each time point by multiplying expression profiles from the SFM condition by the corresponding \( T \) matrix. We quantify the performance of the \( T \) models by comparing the predicted expression profiles with the experimental data and through leave-one-out cross validation (LOOCV). Briefly, in LOOCV one cell line is excluded from the training set, the \( T \) models are constructed using the remaining cell lines, and the SFM expression profile of the removed cell line is used to predict its profile at the next time point. The SVD model performs perfectly on its own training set, but clearly overfits during LOOCV (Fig. 3). The lasso model performs relatively well on both the training set and LOOCV. On average (of all time points and conditions), the lasso models showed 0.95 correlation with the data in full modeling scheme (Fig. 3 and supplemental Fig. S11). In LOOCV analyses, they also performed well, with Pearson correlation coefficient values of 0.902 \( \pm \) 0.005 for IGF1 and 0.895 \( \pm \) 0.006 for Ins stimulated cases (supplemental Fig. S11). The MaxEnt results are not exact for either the training data or LOOCV and are worse than lasso predictions (\( R \approx -0.6 \)). In fact, we notice that the MaxEnt predictions have a negative correlation with experimentally measured values. Based on these performance results, the sparsity of the lasso matrices (supplemental Fig. S9), and the significance of the inferred interactions (see Methods for details), we selected the lasso model as the main model for experimental predictions.

Differences between lasso-inferred interaction networks in IGF1 and Ins stimulated cells can be visually detected, as shown in Fig. 4 and supplemental Fig. S9. Both networks are sparse, and they have very few interactions in common. To verify that our method is robust, we compared the networks obtained through LOOCV with those constructed from the full data set. In the LOOCV, a \( T \) matrix was constructed for excluding each cell line once, giving a total 21 LOOCV models for each time translation. The networks from these models overlap with the full model networks. Although the number of coinciding edges was dependent on cell lines, these results provide additional support that the lasso model inference was robust. The overlap of inferred interactions from the full and LOO models are presented in supplemental Fig. S12.
Differential Signaling of Insulin and IGF1

Fig. 3. Performance analyses of the three models reveal that lasso models are accurate and robust. Three matrices were constructed for full (Fig. 1A) and LOOCV (Fig. 1B) models of each time translation. Correlation plots of data and prediction results from the three models for in-training (left) and leave-one-out cases (right) are shown. Black dots in the left panels represent data versus full model predictions, and black diamonds on the right represent the data versus LOOCV model predictions. The red lines are the least-square fits. The Pearson correlation coefficients are reported on each subpanel as R. Overall, the lasso models outperformed the other two methods. Modeling results of time translation for SFM to 10 min IGF1 stimulation in MCF7 cells shown as an example.

representative LOOCV models constructed without the information for the corresponding cell lines are shown with comparison to full model interaction matrices. The mean Jaccard coefficients for comparison of LOOCV model networks with full model networks are 0.65 (±0.063) and 0.64 (±0.065) for IGF1 and insulin cases, respectively. These numbers mean that out of the interactions inferred for full and LOOCV models, on average 65% of them are found in both of the corresponding time point models and that the lasso models are robust in interaction inference. Additionally, the lasso model matrices are 83% (±2) sparse both in full and LOOCV models.

Lasso-inferred Interactions are Experimentally Validated—We constructed six TT matrices for IGF1 and six matrices for Ins using the mean SFM condition data as X(t) and data from each stimulation time point as X(t) (see Methods). In this setting, the TT matrices reflect the interactions from time zero to the end of hormone stimulation. We used SFM as the starting point for all TT calculations because SFM protein levels can be experimentally controlled, enabling us to validate model predictions.

The interaction values of the TT matrices reflect the effect of synthetically knocking down a protein. If one perturbs the interaction values of a protein (a column in the TT matrix) and predicts how the responses deviate from the data, the results are correlated with the magnitudes of the inferred interaction constant. These observations led to the study of the individual interaction values only. First, the set of nonself-interactions present in TT matrices of both IGF1 and Ins stimulation cases were determined (96, 87, 98, 78, 62, and 85 interactions found for time points 1–6, respectively). The list was filtered to retain only pairs of interactions having either phospho-Akt (pAkt) or phospho-MAPK (pMAPK) as outputs because these are two of the major downstream mediators of IGF1 and Ins signaling. Finally, for each time point the interactions were ranked by differential response to IGF and Ins (i.e. for each interaction in each time point, the quantity \( \frac{\log(TT_{IGF})}{\log(TT_{Ins})} \) was calculated, where \( TT_x \) is the response in condition \( x \)). The depiction of the top exclusive interactions for IGF1 and insulin cases as well as the top differential interactions revealed how complex our study system is (Fig. 4).

Our model identifies a number of previously known interactions. For instance, the magnitudes of the inferred interactions between PKCa - pPKCa and between Rb1 - pRb1 are among the highest strength interactions. These indicate that the level of nonphosphorylated proteins in unstimulated cells influences the level of phosphorylated proteins of the same protein at later times (see supplemental Table S5 for more such interactions). It is also important to note that the existence of these two interactions reflects on the importance of the roles of PKCa and Rb1 in cancer progression and transcriptional regulation, modulated downstream of IGF1 and Ins signaling (63–68). Additionally, the inferred interaction of Rb1 level affecting pRb1 level is the only nonself interaction present in all eight (four IGF1, four Ins) lasso models up to 6 h, pointing out the importance and critical role of Rb1 regulation. Other such interactions are also uncovered by the model, including the levels of phosphorylated Akt (69), EGFR (70, 71), and HER2 (72, 73) depending on the total protein level in SFM condition.

Our method also predicts several interactions (supplemental Table S6) that are neither reported in the literature nor in databases like STRING (74). For example, a STRING search using the 42 gene names (supplemental Table S1) corresponding to the 43 proteins used in this study produces a network of 49 edges based on a confidence cutoff of 0.9 and evidence only from experiments. None of the interactions reported in Table I and supplemental Table S6 are in that list. When the STRING-based network is compared with our lasso models, there are 12 and 10 edges in common for IGF1 and Ins networks, respectively. The probabilities of having at least
these many edges correctly assigned in a random network (calculated via the hypergeometric distribution) are $P_{IGF} = 0.021$ and $P_{Ins} = 0.063$, suggesting that the lasso models capture more known interactions than would be expected at random. The list presented in Table I summarizes results from the lasso models. The first line indicates that phospho-ACC (pACC) levels are predicted to have a greater effect on the IGF1-induced phosphorylation of MAPK than on insulin-induced MAPK phosphorylation after 5 and 10 min of stimulation. The SVD model makes the same prediction, but extends it to 30 min after stimulation (supplemental Table S6). The MaxEnt model also predicts the effect for 5 min of stimulation (supplemental Table S6), but interestingly predicts that the effect is stronger because of insulin stimulation for 10 and 30 min stimulation. Importantly, all three models predict that altering levels of pACC in cells in SFM will impact levels of pMAPK upon IGF1 and Ins stimulation, and that the extent of the effect will depend on which growth factor is introduced. The interaction was also shown to be robust, appearing in 99.96% of networks (out of 40,000) generated for the first two time points for both insulin and IGF1 stimulation in LOOCV of the HCC1569 cell line. It is also important to note that our models do not discriminate between direct and indirect associations, so that the interactions that we observe may be mediated by other cell factors that are not measured in the current data set.

We verified the differential influence of ACC using knock-down experiments. The two cell lines for the knock-down, MCF7 and HCC1569, were chosen because they have high pACC protein levels in the data, and also because both exhibit an expected MAPK response upon hormone stimulation. The immunoblot shown in Fig. 5A and the bar graph in Fig. 5C demonstrate that upon ACC1/2 siRNA knock-down in MCF7 cells, phospho-MAPK levels increased, and that the extent of the increase was different in IGF1- and Ins-stimulated cells. Compared with Scr cells, IGF1 stimulated cells with no active ACC1/2 showed higher MAPK phosphorylation, amounting to a 40% increase ($p < 0.05$) in fold-change level at 5 min. However, a twenty percent decrease in fold-change ($p > 0.05$) was observed in insulin stimulated cells, suggesting a higher proliferative IGF1 response compared with insulin induction. At 10 min, there was an insignificant down-regulation of the pMAPK response in ACC knock-down cells in both cases. Additionally, we showed that the ACC knock-down in
HCC1569 cell line causes a significant change (70% increase in fold-change, \( p < 0.05 \)) in IGF1 stimulated cells compared with an insignificant change (5% increase, \( p > 0.05 \)) seen in Ins stimulated cells at 10 min (supplemental Fig. S13). Indeed, the differences in fold-change of relative pMAPK levels were larger in IGF1 stimulated cells for all time points in HCC1569 cells (supplemental Fig. S13).

Another interesting prediction of our model is the differential effect of E-Cadherin expression levels on pAkt level induced by IGF and Ins. This is the top differential effect candidate for the 30 min time point, and was shown to be present in 97.35% of LOO networks constructed for 10 and 30 min stimulation of both growth factors. Figs. 5B and 5D show increased Akt activation upon E-Cadherin knockdown. The effect was more drastic in response to IGF1 than to Ins. Leading to higher growth signaling in stimulated cells, the loss of E-Cadherin induced an almost 2.5-fold increase in relative pAkt/Akt level in response to both IGF1 (\( p < 0.005 \)) and Ins (\( p < 0.01 \)) at 10 min. At 5 min, the change was significant again in both cases (\( p < 0.01 \) and \( p < 0.005 \), respectively). The magnitude of the increase in fold-change was twice in IGF1 stimulated cells compared with insulin induced cells both at 5 and 10 min.

**DISCUSSION**

In the present work, we analyzed temporal proteomic data to identify differences in the insulin and IGF1 signaling pathways in a large panel of human breast cancer cell lines. The data consist of protein expression levels generated by RPPA, collected at different times with IGF1 and Ins stimulation conditions. The transformation matrices obtained by the lasso approach differ significantly from matrices generated using randomized data in the number of interactions (supplemental Figs. S3 and S4), and in interaction magnitudes (supplemental Fig. S5). We predicted that knocking down ACC and E-cadherin would produce differential effects in Ins- and IGF1-stimulated cells, and we verified the predictions experimentally.

RPPA acquisition and analysis is still a growing area of research. Studies utilizing RPPA data have employed a diverse range of data (pre)processing and benchmarking methods, but no single protocol for processing RPPA data has yet been universally accepted (32–34, 75). Although inclusion of
RPPA is still lacking, there are ongoing efforts to standardize and coordinate dissemination of different data types (76). Here, we mean-centered and standardized the protein arrays before model construction. This pre-processing enabled us to study the relative shapes of the time-course expression curves rather than of the amplitudes. Each sub-matrix of 43 proteins in 21 cell lines was processed separately, and three different models of time-translation were constructed for each pair (Fig. 1).

The lasso model TT matrices constructed for each time point pair were sparse (supplemental Fig. S9). They performed overall best among the three modeling schemes explored (SVD, MaxEnt, and lasso). The lasso algorithm provides a means of selecting a few good descriptor weights given input data, to recapture the output with minimal deviation (58, 60, 77–79). It was shown in Fig. 3 that the lasso models perform equally well in leave-one-out tests as in training-set tests. When compared with the other two models, the lasso model selected enough good descriptors during model construction to enable better predictions from unseen input data. The lasso models also contain fewer nonzero interactions than the other models, easing the burden of identifying biologically relevant interactions. This lasso procedure follows a gradient descent method that depends on a random seed. To minimize the influence of random effects on our results, a large number of matrices were generated and one representative canonical lasso matrix was selected based on the number of times that each of its elements is nonzero. Selection of this ensemble median network as the model representative is also useful in other applications, such as reconstruction of gene regulatory networks from microarray data (80, 81).

The covariance-matrix derived models (SVD) are analytical solutions of the training problem and suffer from overfitting. These models can recapture the response exactly for the training data, whereas outputs for nontraining-set inputs deviate from the real response considerably. The overfitting can be relaxed by invoking maximum entropy (MaxEnt models), but this gives rise to another problem: Because the number of proteins is larger than the number of cell lines the MaxEnt method does not exactly reproduce the training set data. Instead of predicting the training data exactly and the test data poorly, it generates mildly erroneous predictions for all data. Interestingly, we find that the data agree more closely with the inverse of the MaxEnt predictions than with the predictions themselves. This discrepancy between measured and modeled values occurs only when the number of proteins is greater than the number of conditions. If we use 20 proteins instead of 43, the data agree better with the MaxEnt solution than with its inverse. The MaxEnt matrix has been shown in earlier works to be more stable for systems for inputs with random noise (48). The lasso models outperformed maximum entropy models here, suggesting that the data are not drawn from a maximum entropy distribution.

Our model predictions were validated in multiple ways. First, existence of post-translational modifications in the translation matrices, i.e. effects of nonphosphorylated forms of PKCζ and Rb1 on the levels of respective phosphorylated states, are important measures of validity of the models. Existence of such interactions over the time-course persistently shows that the lasso algorithm selects meaningful, nonrandom descriptors. Further validation comes from the presence of known interactions between kinase and substrate. Examples from the lasso model include the level of p70S6K affecting the level of phospho-S6 (82) and the predicted dependence of TSC2 phosphorylation level on the level of Akt present. TSC2 is phosphorylated in response to activation of Akt, leading to activation of mTORC1 complex (83). A third tier of validation is the set of results obtained from the in vitro knock-down experiments. The effects of ACC knock-down on pMAPK level and of E-Cadherin knock-down on pAkt level in response to IGFI and Ins show that the model predictions are accurate. The fact that the aforementioned experimental candidates show differential results in IGFI and Ins stimulation conditions validates the overall approach for elucidating novel interactions downstream of the two hormones.

ACC is involved in lipid metabolism, specifically in conversion of acetyl-CoA to malonyl-CoA. The latter is the building block of fatty acids and there is evidence tying ACC activity to the energetic state of the tumor microenvironment and cell survival (84, 85). Low energy states (high AMP/ATP ratio conditions) promote higher AMPK activity, inhibiting ACC1/2 by phosphorylation. The inhibition of ACC in turn leads to increased fatty acid oxidation and decreased lipid biosynthesis (85). It has been suggested that high levels of inactive pACC sustain NADPH (nicotinamide adenine dinucleotide phosphate), a reducing agent for biosynthetic reactions, at a high level, providing a better environment for tumor cell growth under stress. It has also been shown that the diabetic drug metformin activates AMPK, leading to an increase in pACC levels and consequently to higher insulin signaling and reduced insulin resistance (84). Although there is evidence linking pACC to cell growth and insulin signaling, we found no previously reported link between pACC and pMAPK. Our results are consistent with the previous results such that a decrease in ACC1/2 level promotes cell survival, although through an additional mechanism of action. This differential effect in response to IGFI or Ins requires further study.

The following may be an example of our model identifying an indirect effect between proteins including E-Cadherin and Akt. There are controversial results reported so far, some stating that the presence of E-Cadherin junctions between cells recruits PI3K to the cell membrane, resulting in a higher activation of Akt in normal kidney and ovarian cancer cells (86, 87). Lau and colleagues, on the other hand, stated that endogenous E-Cadherin signaling negatively regulates Akt activation by sequestering beta-catenin leading to increased PTEN transcription and decreased Akt phosphorylation (88). Loss of E-Cadherin then releases beta-catenin for translocation into the nucleus leading to inhibition of PTEN transcrip-
tion and thus up-regulation of Akt signaling. Complementarily, previous investigations showed that upon IGF1 treatment, increased pIGF1R and pAkt levels reduced E-Cadherin concentration and functioning (89–92). It was also shown that high cell-cell contact formation by E-Cadherin disrupted high affinity IGF1 binding to its receptor, resulting in lesser downstream activation (93, 94). Collectively, our lasso model prediction clearly has foundations in the literature but the differential effect in response to IGF1 and insulin has not been studied. Our results support the notion that loss of E-Cadherin induces pAkt and also coincide with the earlier works showing a relationship between IGF1 signaling and EMT (Epithelial-to-Mesenchymal Transition) through regulation of E-cadherin.

Fig. 6 summarizes our hypothesized mechanisms of actions for ACC and E-Cadherin within the canonical cascades of IGF and Ins signaling.

Another interesting prediction of our lasso models was the effect of ATM on both phospho-forms of Akt (Table I). Earlier studies showed that ATM mediates full activation and downstream signaling of Akt under insulin and IGF1 stimulation in other tissue types (95–97). Our models are of note to infer edges from ATM to both pAkt proteins, where full activation of Akt is accomplished by dual phosphorylation of these sites. Further study of that interaction might spearhead new combination therapies with IGF1R inhibitors in breast cancer tumors. Indeed, very recently, researchers have shown that presence of kinase-domain mutations in ATM is a biomarker for Topo-isomerase I inhibitor therapy (98).

The study of breast cancer prevention, diagnosis, and therapy is imperative. However, compensation and activating feedback mechanisms collectively depress the efficacy of anti-IGF1R/InsR therapies (11). Only after discerning the dual IGF1R/InsR system dynamics and novel components, will we be able to start to direct patients with suitable tumor subclasses toward more efficient personalized clinical interventions. The computationally scalable reverse-engineered models of cellular networks introduced here provided us with a framework to elucidate experimental targets of pharmacological importance in a cost effective way.

* This work was supported by the National Center for Advancing Translational Sciences at the National Institutes of Health under Award Number U54TR0005032, by the Commonwealth of PA under Award Number 4100062224, by NCI under Award Number P50CA058183 and R01CA94118, and by NIH under Award Number T32 GM008424. RPPA Core Facility, where the initial data set was generated, is funded by NCI Award Number CA16672. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
Differential Signaling of Insulin and IGF1

10. Pollak, M. (2008) Insulin and insulin-like growth factor signalling: a therapeutic approach for breast cancer. J. Mammary Gland Biol. Neoplasia 11, 27–39
21. Niederst, M. J., and Engelman, J. A. (2013) Bypass mechanisms of resistance to receptor tyrosine kinase inhibition in lung cancer. Sci. Signal. 6, re6
22. Lu, Y., Zi, X., Zhao, Y., Mascarenhas, D., and Pollak, M. (2001) Insulin-like growth factor-I receptor signaling and resistance to trastuzumab (Herceptin). J. Natl. Cancer Inst. 93, 1852–1857
23. Buck, E., Gokhale, R. C., Kousak, S., Brown, E., Eyzaguirre, A., Tao, N., Rosenfeld-Franklin, M., Lerner, L., Chiu, M. I., Wild, R., Epstein, D., Pachter, J. A., and Migliarese, M. R. (2010) Compensatory insulin receptor (IR) activation on inhibition of insulin-like growth factor-1 receptor (IGF-1R): rationale for cotargeting IGF-1R and IR in cancer. Mol. Cancer Ther. 9, 2652–2664
24. Lemmon, M. A., and Schlessinger, J. (2010) Cell signaling by receptor tyrosine kinases. Cell 141, 1134–1155
25. Cancer Genome Atlas, N. (2012) Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70
26. Chakravarti, A., Leoffler, J. S., and Dyson, N. J. (2002) Insulin-like growth factor-I receptor mediates resistance to anti-epidermal growth factor receptor therapy in primary human glioblastoma cells through continued activation of phosphoinositol 3-kinase signaling. Cancer Res. 62, 200–207
27. Jones, R. V., Goddard, D., Yang, L. W., and Bahar, I. (2004) Structural basis for bispecific antibody overcomes network adaptations that limit activity of IGF-IR inhibitors. Mol. Cancer Ther. 13, 410–425
28. Senutovitch, N., Vernetti, L., Boltz, R., DeBiasio, R., Gough, A., and Taylor, D. L. (2015) Fluorescent protein biosensors applied to microphysiological systems. Exp. Biol. Med. 240, 795–808
29. Tibes, R., Qi, Y., Lu, Y., Hennessey, B., Andreuff, M., Mills, G. B., and Kornblau, S. M. (2006) Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Mol. Cancer Ther. 5, 2512–2521
30. Daemen, A., Griffith, O. L., Heiser, L. M., Wang, N. J., Enache, O. M., Sanborn, Z., Pepin, F., Durinc, S., Korkola, J. E., Griffith, M., Hur, J. S., Hu, S., Chung, J., Cope, C., Feckler, M., Cabrales, D., Seth, P., Sukhatme, V. P., Jakubowicz, L., Lu, Y., Mills, G. B., Cho, R. J., Collisson, E. A., van’t Veer, L. J., Spellman, P. T., and Gray, J. W. (2013) Modeling precision treatment of breast cancer. Genome Biol. 14, R110
31. Toettcher, J. E., Weiser, O. D., and Lim, W. A. (2013) Using optogenetics to interrogate the dynamic control of signal transmission by the Ras/Erk module. Cell 155, 1422–1434
32. Iademarco, S., Lu, Y., Morales, F. C., Mills, G. B., and Ram, P. T. (2010) Identification of optimal drug combinations targeting cellular networks: integrating phospho-proteomics and computational network analysis. Cancer Res. 70, 6704–6714
33. Bar-Joseph, Z., Gitter, A., and Simon, I. (2012) Studying and modelling dynamic biological processes using time-series gene expression data. Nat. Rev. Genet. 13, 552–564
34. Jin, L., Luo, X. Y., Su, W. Y., Zhao, X. L., Yuan, M. O., Han, L. Z., Zhao, X., Chen, Y. D., and Rao, S. Q. (2014) Pathway-based analysis tools for complex diseases: a review. Genomics, Proteomics Bioinformatics 12, 210–220
35. Bahar, I., Atlign, A. R., and Erman, B. (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des. 2, 173–181
36. Chennubhotla, C., and Bahar, I. (2007) Signal propagation in proteins and relation to equilibrium fluctuations. PLoS Computational Biol. 3, 1716–1726
37. Chennubhotla, C., Rader, A. J., Yang, L. W., and Bahar, I. (2005) Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies. Phys. Biol. 2, S173–S180
38. Wang, G. (2010) Singularity analysis of the AKT signaling pathway reveals
