Published in final edited form as:

Mol Biosyst. 2014 March 4; 10(3): 605–612. doi:10.1039/c3mb70403c.

Correlating phosphoproteomic signaling to castration resistant prostate cancer survival through regression analysis

Reynald Lescarbeau^a and David L Kaplan^a

^a4 Colby St., Medford, MA, USA.

Abstract

Prostate cancer most commonly presents as initially castration dependent, however in a minority of patients the disease will progress to a state of castration resistant. Here, approaches to correlating alterations in the phosphoproteome to androgen independent cell survival in the LNCaP, PC3, and MDa-PCa-2b cells lines are discussed. The performance of the regression techniques multiple linear, ridge, principal component, and partial least squares regression are compared. The predictive performance of these algorithms over randomized data sets and using the Akaike Information Criterion is explored, and principal component and partial least squares regression are found to outperform other regression approaches. The effect of altering the number of features versus observations on the R^2 value and predictive performance is also examined using partial least squares regression model. Utilizing these approaches “drivers” of castration resistant disease can be identified whose modulation alters phenotypic outcomes. This data provides an empirical comparison of the various considerations when statistically analyzing phosphorylation data with the aim of correlation to phenotypic outcomes.

Introduction

Upon initial presentation prostate cancer is generally androgen dependent, however following treatment to reduce androgen levels, prostate cancer can develop to become castration resistant and prognoses are poor [1, 2]. Similar to many other cancers, mutations and activation of numerous pathways has been described as being able to enable castration resistant disease progression. Major pathways and mechanisms which have been described as being able to drive castration resistance include intracrine synthesis of androgen, activation of PI3K, β -catenin, MAPK, Stat3, or NF- κ B amongst others [3-7].

The goal of the present work is to illustrate and compare statistical algorithms for relating phenotypic alterations to cell signaling pathways in castration resistant prostate cancer. The levels of pathway activation, determined via measuring the phosphorylation level of numerous kinases, are regressed against the survival levels of various prostate cancer cell lines under the same treatments. These treatments include the growth factors EGF and IGF, the cytokines IL6 and TNF α , dihydrotestosterone (DHT) an androgen receptor agonist, and the chemotherapeutic docetaxel. These treatments are combined with five targeted kinase inhibitors in LNCaP cells to examine the effect of these drug treatments on modulating cell signaling, and to create a more diverse, perturbation rich, data set.

Previous work has utilized regression analysis to correlate various phenotypic outcomes to alterations in cell signaling such as migration and cell proliferation [8, 9]. In these and other

works, dimensionality reduction techniques have also proven to be useful in collapsing numerous features (i.e. phosphoproteins) into latent variables which can then be used for regression analysis (partial least squares regression) or data visualization (principal component analysis) [8]. Regression approaches based on this method as well as multiple linear regression using ordinary least squares and regularization regression (ridge regression) are compared [10, 11].

The use of these quantitative techniques is described in terms of the new application of predicting prostate cancer castration resistant growth in three cell lines with varying degrees of sensitivity to androgen depleted conditions. Additionally, the accuracy of the different regression algorithms is compared via a permuted benchmark test and the Akaike Information Criterion which takes into account the model degrees of freedom. Although prostate cancer castration resistant growth is specifically examined, to our knowledge, this work represents the first side-by-side comparison of these common regression algorithms as applied to a phosphoproteomic data set with the goal of predicting cell fate.

Methods

Cell culture

LNCaP, MDA-PCa-2b, and PC3 cell lines were acquired from ATCC (Manassas, VA, USA). PC3 and LNCaP cells lines were cultured in 10% fetal bovine serum (FBS), RPMI 1640, and 1% antibiotic-antimycotic. The MDA-PCa-2b cell line was cultured in BRFF-HPC1 media purchased from AthenaES (Baltimore, MD, USA) supplemented with 20% FBS. Dihydrotestosterone was acquired from Sigma-Aldrich (St. Louis, MO, USA). Androgen depleted media consisted of 10% charcoal stripped FBS with phenol red free RPMI 1640. Docetaxel was acquired from Sigma-Aldrich. Temsirolimus and SB202190 were purchased from Selleckchem (Houston, TX, USA). All other inhibitors were purchased from EMD Millipore (Billerica, MA, USA). Unless otherwise stated all other cell culture reagents were acquired from Invitrogen (Grand Island, NY, USA).

Measuring cellular response to treatment

Relative cell viability was assessed using an MTT ((3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) assay acquired from Invitrogen. Our lab has previously determined that MTT correlates to relative cell number as confirmed via DNA quantification and manual cell counting [12]. All three cell lines were plated at a concentration of 5,000 cells/cm² in a 24 well plate in their respective growth media. The cells were allowed to adhere for 24 hours. The media was then changed to androgen depleted media which the cells were cultured in for an additional 72 hours. Finally, relative cell viability was determined using an MTT assay according to the manufacturer's instructions. Targeted kinase inhibitors were used at the following concentrations: LY294002 at 7 μ m, U0126 at 325 nm, Wedelolactone at 10 μ m, Temsirolimus at 50 nm, and SB202190 at 500 nm. Additionally, the total protein amount of biological replicates from each cell type was measured using a Bicinchoninic assay purchased from Thermo Scientific (Rockford, IL, USA). After measuring cell survival with a MTT assay the results were normalized to total protein measured to account for variations in cell size between the cell lines.

Cell lysates

Each prostate cancer cell line was plated to six well plates at a density of 5,000 cells/cm² in their respective growth media and allowed to adhere for 24 hours. After 24 hours cells were treated with androgen depleted media supplemented with the appropriate treatment. For studies involving the use of inhibitors on LNCaP cells, the cells were first pretreated for 30 minutes with the inhibitor before additional treatments were added. Following the

appropriate amount of time (30 minutes, 4 hours, or 24 hours) the media was removed and the cells were lysed.

384 well phosphoproteomic ELISA assay

R&D Systems (Minneapolis, MN, USA) DuoSet ELISA kits were used to quantify the amount of phosphoprotein present in each sample in 384 well plates. Lysates were processed and the assays performed according to manufacturer's instructions with the exception that they were modified for a 384 well format. This modification involved simply reducing the total volume of reagents recommended per 96 well by one fourth. A Bicinchoninic acid assay was performed on each lysate and the phosphoprotein levels were normalized to the total amount of protein per lysate.

Data analysis

Following data acquisition, calibration to the ELISA standard curve, and normalization to total protein content, both protein (X matrix) and survival data (Y matrix) was imported into Matlab (The Mathworks, Natick, MA) where analysis was performed. The data was arranged such that each column of the X matrix represented a phosphoprotein at a specific time (8 phosphoproteins X 3 time points = 24 columns). The rows represent the cell treatments with the values in the X matrix corresponding to phosphorylation levels and the Y matrix equals to relative cell survival in response to that treatment. Partial least squares regression was performed using the native *plsregress* function which is implemented via the SIMPLS algorithm while principal component regression and analysis was performed using the *princomp* function with the stated number of principal components. Ridge regression was performed utilizing the *ridge* function with the optimal theta determined by the Hoerl-Kennard method.

Results

Measuring castration resistant growth in response to treatment

When examining castration resistant growth, our own studies have indicated that high cell densities can cause growth arrest cells to be less affected by androgen depleted conditions so cell densities are kept at 30-50%. The androgen depleted media either contained a DMSO vehicle control, EGF, IGF1, IL6, TNF α , DHT or Docetaxel. These treatments were additionally combined with LY294002 (PI3Ki), U0126 (MEKi), wedelactone (IKK α / β i), temsirolimus (mTORi), and SB202190 (p38i) when treating the androgen dependent cell line LNCaP cells. The basal level of castration resistance of each cell line to androgen depleted media with various treatments can be seen (Figure 1A), as well as the phosphosites measured (Table 1). Additionally, the relative effect of each treatment on cell survival is illustrated.

Quantifying cell lysates for phosphoprotein levels

The protein lysates were collected at 30 minutes, 4 hours, and 24 hours following the given treatment, and the total amount of protein per lysate was quantified. Ninety six well ELISA assay kits were then modified for a 384 well format by scaling the volume of reagents used down to one fourth the original volume. Standard curves were generated with positive control samples to determine the absolute amount of phosphoprotein contained in each sample (Supp. Fig. 1). Due to limitations in performing large numbers of replicates each treatment was performed with one biological duplicate. To ensure good repeatability the absolute percent difference between the two identical samples was determined, and the average value calculated per cell line and per phosphosite (Table 2).

Finally, the lysates were taken at three different time points (30 minutes, 4 hours, and 24 hours) to capture the time dependent alterations in signaling which occur. Generally, phosphoproteomic alterations occur over the course of minutes in response to changes in microenvironment. Negative feedback and transcriptional changes may then cause longer term alterations changes in signaling over the course of hours before the phosphosites reach a steady state value. For this reason three time points were chosen in this study.

Data processing methods

Following calculation of the total mass of phosphoprotein present, the duplicate values were averaged and arranged in a matrix with rows representing the treatment (cell line, growth factor, and possibly drug) and the columns representing the phosphosite and time point. When performing regression analysis we normalized all protein values to their mean due to the fact that different phosphoproteins had innately different amounts of protein present, rather than preserving the innate differences in amounts by taking the log of the data. This mean centered data was then unit variance scaled (Z-score) to express it in standard deviations from the mean. This mean centering and variance scaling was performed on the same phosphoprotein across all time points.

Following normalization and preprocessing of the data, hierarchical clustering using a Euclidean distance metric was used to determine the relationship between the data (Figure 1B) as previously discussed [13]. Cell line data clustered most strongly by cell line and then inhibitor treatment, indicating these as having the maximum amount of variation on cells. Principal component analysis (PCA) was also performed to examine localization differences in the treatment groups in principal component space. Principal component analysis when applied to phosphoprotein data collapses multicollinear features (correlation in features) into principal components. These principal components can then be plotted in 2 or 3 dimensions and systematic variations due to particular treatments observed. Here, the data was plotted on 2D biplot using the first and second principal component (Figure 2A). From this data it can be seen that PCA separated the data largely based on cell line which indicates there are likely innate variances between the cell lines which dominate the effects of any ligand or drug treatments. The breakdown of the phosphoprotein at each time point that contributed to each principal component can also be seen (Figure 2B and 2C). The first principal component has a substantial number of phosphoproteins which are part of PI3K signaling weighted positively (Akt, GSK3, and RPS6). In our previous work these three phosphoproteins were found to positively correlated (data not shown). All three time points for p38 and JNK were negatively weighted in the first principal component. In the second principal component the most notable patterns were strong positive weightings on JNK and Stat3.

Performing regression analysis on data sets

As previously discussed, various regression analyses can be performed in an attempt to correlate variations in phosphoproteins to cell fate outcomes. Here, the differences between multiple linear regression (ordinary least squares), partial least squares (three principal components), principal component regression (ten principal components), and ridge regression are examined (Figure 3). For partial least squares regression it was determined with the data set presented here that three principal components was an optimal stopping point [14]. With principal component regression more principal components were included because these principal components only model the variation in X, without respect to Y. For this reason principal components with lower eigenvalues may contribute substantially to increasing the accuracy of the regression. Additionally, the resulting R^2 values calculated for each of these regressions is depicted (Table 3). The second and sixth principal components

when performing principal component regression were found to most substantially contribute to the overall model (Figure 4).

Comparing regression model performance

In order to determine the significance of the model as compared to the null hypothesis, the response labels were permuted with respect to the phosphoprotein data. From this randomized regression model an R^2 value is determined. This is repeated for a given number of times to get a distribution of R^2 values corresponding to randomized data. Assuming normality, this randomized data is then compared to the original R^2 value from the accurate model. This was done for all four regression methods used and the mean R^2 from the randomized method, as well the standard deviations are shown. Using this data we calculated the number of standard deviations above the randomized mean R^2 that the actual R^2 value for the full model was (Table 3). It can be seen that principal component regression appeared to have the best performance, however this was with ten principal components.

The regression coefficients which were calculated from these regression analyses were plotted side by side for comparison (Figure 5A). Hierarchical clustering analysis was also applied to these regression coefficients to determine their similarity to each other (Figure 5B). It can be seen that principal component regression and partial least squares regression returned similar regression coefficients, and that ridge regression and multiple linear regression returned similar regression coefficients. Additionally, principal component and partial least squares regression returned similar R^2 values (0.5819 vs. 0.5961), as did ridge and multiple linear regression (0.7276 vs. 0.7515).

To further compare the model performance the Akaike Information Criterion (AIC) was calculated which takes into account the sample size, residual sum of squares, and degrees of freedom. A lower AIC indicates a better model, and the AIC value is penalized for an increasing number of degrees of freedom and higher residual sum of squares while it rewards increased sample size. The AIC formula is:

$$AIC = \log \left(\frac{RSS}{n} \right) + 2k$$

Where RSS is the residual sum of squares, n is the sample size, and k is the degrees of freedom. Determining the degrees for freedom for PCR, MLR, and ridge regression is straightforward, however to calculate the degrees of freedom for PLS regression a newly described approach was used since the number of principal components themselves has been shown to underestimate the degrees of freedom [15]. Upon examining the calculated AIC values they were found to generally agree with the comparison of R^2 values to randomized regressions. The PC regression had the lowest AIC (63.8) followed by PLS regression (69.5) which is similar to what was seen with the calculated z-score of the calculated R^2 value as compared to the random R^2 value distribution (Table 3). Additionally, ML and ridge regression had higher AIC values and correspondingly lower z-scores.

Model performance based on number of observations

A pertinent observation surrounding the correlation of signaling pathways to cell fate is the number of treatments included versus the number of features can strongly affect the outcomes. In particular, a large number of features as compared to observations is an area that partial least squares regression attempts to address. To quantitatively study this effect we performed partial least squares regression (3 principal components) on randomly selected subset of the total data, iteratively, and compared those R^2 values to the same randomly

selected data with the Y values (survival) randomized with respect to the X data (phosphoprotein data) (Figure 6A). Based on this analysis we further examined the number of random, or nonsense, regressions which produced R^2 values high than the average R^2 value for that number of selected observations. The percentage which were higher can be seen (Figure 6B). When 5 or 7 of the 63 rows are selected the “true” partial least squares regression produces an R^2 value lower than the same data randomized to each other in a majority of cases. The number of randomized regressions which exceed the mean of “true” PLS regression does not go below 5% of until 25 or more observations are included.

Discussion

Currently, a major area of research in systems biology is the identification of a predictive set of biomarkers to determine phenotypic or disease outcomes. Given that many biological data sets are multicollinear (contain features which are linear combinations of each other), regression approaches which collapse the data set or perform feature selection are likely to be more stable than multiple linear regression. Here, we rigorously examine the performance of several regression methods for multivariate biomarker discovery.

A quantitative comparison of various regression algorithms can be seen with principal component and partial least squares regression having the highest accuracy over other regression techniques as compared to randomized controls (Table 3). This was confirmed via a comparison to both randomized data, as well as the use of the Akaike Information Criterion. It is important to note that if the number of degrees of freedom used in PLS regression was equal the number of principal components used (which was three) then PLS regression would have given the best AIC. Instead a more accurate approach which takes into account the multicollinearity of the data set was employed [17]. Finally, it was seen that principal component regression and partial least squares regression had very similar regression coefficients which is as expected given their similarity in methodology, while ML and ridge regression were more similar in their regression coefficients.

It was additionally observed that for partial least squares, while the R^2 value decreases as the ratio of observations to features increases, the statistical validity, and likely out-of-sample performance increases as well. With this data set, there were less than 5% of the randomized regressions having a greater R^2 value than the average true R^2 value until the number of observations equaled the number of features (24×24 matrix). This likely varies substantially with the data set, however it indicates that regressions performed with fewer observations may produce high R^2 but lack statistical validity.

Here, various methods for analyzing phosphoprotein data and correlation to phenotypic outcomes are examined as well as an empirical comparison of various regression algorithms applied to this data. It was discovered the PC and PLS regression performed very similarly, producing similar R^2 values, regression coefficients, and AIC values. It was additionally noted that these regression approaches which employ dimensionality reduction performed substantially better than standard multiple linear regression or ridge regression. The use of these data driven, statistical approaches for correlating cell signaling to cell fate, offers the benefits of an unbiased examination of multiple pathways simultaneously giving a multivariate view of the drivers of cell fate. In future work, these approaches could be applied to patient specific phosphoprotein data to potentially predict the level of hormone resistance of a patient's disease and indicate pathways which are driving resistance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the NIH (P41 EB002520) for support of this work.

References

1. De Bono JS, Logothetis CJ, Molina A, Fizazi K, North S, Chu L L, et al. *N. Engl. J. Med.* 2011; 364:1995–2005. [PubMed: 21612468]
2. Scher H, Beer TM, Higano CS, Anand A, Taplin ME, Efstathiou E, et al. *Lancet.* 2010; 375:1437–46. [PubMed: 20398925]
3. Mostaghel EA, Nelson PS. *J. Clin. Endocrinol. Metab.* 2010; 22:243–258.
4. Wang G, Wang J, Sadar MD. *Cancer Research.* 2008; 68:9918–9927. [PubMed: 19047173]
5. Zhu M, Kyprianou N. *Endocrine-Related Cancer.* 2008; 15:841–849. [PubMed: 18667687]
6. Mukherjee R, McGuninness DH, McCall P, Underwood MA, Seywright M, Orange C, Edwards J. *British Journal of Cancer.* 2011; 104:1920–28. [PubMed: 21559022]
7. Nadiminty N, Lou W, Sun M, Chen J, Yue J, Kung HJ, Evans CP, Zhou Q, Gao AC. *Cancer Research.* 2011; 70:3309–3319. [PubMed: 20388792]
8. Kumar N, Wolf-Yadlin A, White F, Lauffenburger DA. *PLoS Computational Biology.* 2007; 3:e4. [PubMed: 17206861]
9. Fitzgerald JB, Schoeberl B, Nielsen NB, Sorger PK. *Nature Chemical Biology.* 2006; 2:458–466.
10. Alexopoulos LG, Saez-Rodriguez J, Cosgrove BD, Lauffenburger DA, Sorger PA. *Molecular and Cellular Proteomics.* 2010; 9:1849–1865. [PubMed: 20460255]
11. Wu CJ, Cai T, Rikova K, Merberg D, Kasif S, Steffen M. *PLoS ONE.* 2009; 4:e7994. [PubMed: 19946374]
12. Lescarbeau RM, Seib FP, Prewitz M, Werner C, Kaplan DL. *PLoS ONE.* 2012; 7:e40372. [PubMed: 22870197]
13. Lescarbeau, RM.; Kaplan, DL. *Molecular and Cellular Proteomics.* Submitted
14. Peres-Neto PR, Jackson DA, Somers KM. *Computational Statistics and Data Analysis.* 2005; 49:974–997.
15. Krämer N, Sugiyama M. *J. Am. Stat. Assoc.* 2011; 106:697–705.
16. Kinkade CW, Castillo-Martin M, Puzio-Kuter A, Yan J, Foster TH, Gao H, Sun Y, Ouyang X, Gerald WL, Cordon-Cardo C, Abate-Shen C. *J. Clin. Invest.* 2008; 118:3051–3064. [PubMed: 18725989]
17. Floc'h N, Kinkade CW, Kobayashi T, Aytes A, Lefebvre C, Mitrofanova A, Cardiff RD, Califano A, M Shen M, Abate-Shen C. *Cancer Research.* 2012; 72:4483–4493. [PubMed: 22815528]
18. Schutz S, Schrader AJ, Zengerling F, Genze F, Cronauer, Schrader M. *PLoS ONE.* 2011; 6:e25341. [PubMed: 21980429]
19. Gan L, Wang J, Xu H, Yang X. *Prostate.* 2011; 71:1158–1166. [PubMed: 21656826]
20. Hubner A, Mulholland DJ, Standen CL, Karasarides M, Cavanagh-Kyros J, Barrett T, Chi H, Greiner DL, Tournier C, Sawyers CL, Flavell RA, Wu H, Davis RJ. *Proc. Natl. Acad. Sci. U. S. A.* 2012; 109:12046–12051. [PubMed: 22753496]
21. Zoubeidi A, Zardan A, Wiedmann RM, Locke J, Beraldi E, Fazli L, Gleave ME. *Cancer Research.* 2010; 70:2307–2317. [PubMed: 20197463]
22. Lee SO, Lou W, Johnson CS, Trump DL, Gao AC. *Prostate.* 2010; 60:178–186. [PubMed: 15176047]

Figure 1.

Overview of the prostate cancer cell survival and phosphoproteomic signaling data. A) The relative cell survival for PC3, LNCaP, and MDA-PCa-2b cell lines under androgen depleted conditions (control), as well as androgen depleted media supplemented with EGF, IGF1, IL6, TNFα, DHT, or Docetaxel. Cell survival was measured after 72 hours using an MTT assay. Cells were plated at the same amount of protein initially because of cell size differences between cell lines. B) Heat map of phosphoprotein levels (left heatmap) and survival levels (column matrix) which were mean centered and variance scaled. Hierarchical clustering was applied in both the x and y direction of the phosphoprotein data.

Figure 2. Principal component analysis applied to the collected phosphorylation data. A) A scatter plot illustrating the weightings on the treated groups from the first two principal components. B) The weights on the first principal component which were sorted in increasing order. C) The weights on the second principal component which were sorted in increasing order.

Figure 3. Predicted versus measured survival data for various regression algorithms. A) Predicted versus measured survival as determined with multiple linear regression. B) Predicted versus measured survival as determined with partial least squares regression with three principal components. C) Predicted versus measured survival as determined with ridge regression. D) Predicted versus measured survival as determined with principal component regression with ten principal components.

Figure 4.

The R-squared value as more principal components are added to principal component regression (red squared), and the increase in R-squared value for each principal component added (blue bar).

Figure 5. Analysis and comparison of the regression coefficients. A) The regression coefficients (Betas) for each regression algorithm on each phosphoprotein at the three time points. B) Hierarchical clustering applied to the calculated betas for each of the regression methods.

Figure 6.

The effect of the ratio of features versus observations when performing partial least squares regression. A) N number of observations (rows) were randomly selected from the full data matrix depicted in Figure 1B consisting of three cell lines treated with six treatments as well as LNCaP cells treated in combination with five targeted inhibitors. These randomly selected rows and corresponding survival values were then used to perform partial least squares regression with three principal components and then r-squared value calculated. This was repeated 1000 times for each number of observations. The mean and standard deviation of these r-squared was then plotted (blue bars). The final blue bar is the full regression model performed once ($R^2=0.5961$). A simple approach was then applied, however after randomly selecting n number of rows the X phosphoprotein was randomized with respect to the Y survival values and partial least squares regression was performed. This was repeated 1000 times, the r-squared values averaged, and the standard deviation taken (red bars). B) To compare the average calculated r-squared value for n observations as compared to randomized, “nonsense” regression we determined the percentage of the 1000 randomized regressions which calculated a r-squared value greater than the mean of a real regression (blue bar).

Table 1

Phosphosites measured

Erk1: T202/Y204 [18] Erk2: T185/Y187
Akt1: S473 [18, 19] Akt2: S474 Akt3: S472
RPS6: S235/S236 [19]
GSK3 α : S21 [20] GSK3 β : S9
p38 δ : T180/Y182 [21]
JNK1 and JNK2: T183/Y185 [22] JNK3: T221/Y223
HSP27: S78/S82 [23]
Stat3: Y705 [24]

Table 2

Absolute percent error between biological duplicates averaged across phosphoprotein and cel line.

	HSP27	Erk	Stat3	p38	Akt	RPS6	GSK3β	JNK	Mean
LNCaP	14.0	19.1	22.0	23.7	15.9	30.0	34.1	12.3	21.4
PC3	13.8	10.6	19.9	7.8	6.2	24.6	18.2	26.3	15.9
MDAPCa2b	12.3	5.3	11.4	25.6	12.8	17.4	21.7	8.8	14.4
Mean	13.4	11.7	17.8	19.0	11.6	24.0	24.7	15.8	20.4

Table 3

Calculated R-squared values versus randomized R-squared for each regression method. Randomized mean and standard deviations were calculated using 1000 permutations and Z-scores calculated.

	R2	Randomized Mean R2	Randomized Std Dev	Z-score	AIC
ML	0.7515	0.3640	0.0915	4.2364	86.0176
Ridge	0.7276	0.3174	0.0977	4.1987	84.7299
PC	0.5819	0.1606	0.0669	6.2952	63.8291
PLS	0.5961	0.2392	0.0718	4.9734	69.5037