Predictive Modeling Case Study on Statistical Difficulties in Combining Genomic and Proteomic Data

Dr. P. Umamaheswari[1]; N. Purusothaman[2]; P. Gayathridevi[3]

[1]Assistant Professor of Statistics, Sona College of Arts & Science,
[2]Department of Biotechnology Sona College of Arts & Science
[3]Department of Biotechnology Sona College of Arts & Science

Publication Date: 2026/01/15

Abstract: A key challenge in modern biology is integrating different types of molecular data. This study examines the specific relationship between gene copy number and protein expression levels. Using data from the Cancer Cell Line Encyclopedia (CCLE), we find that this relationship varies significantly by gene. For the MYC oncogene, copy number strongly predicts protein levels ($R^2 = 0.37$), indicating that more gene copies generally lead to more protein. However, for the TP53 tumor suppressor, copy number poorly predicts protein abundance ($R^2 = 0.08$), suggesting that other regulatory mechanisms dominate. These results show that simple statistical models are often insufficient for biological data, and more advanced approaches are needed to understand complex gene-protein relationships.

Keywords: *Data Integration, Genomics, Proteomics, Copy Number Variation, MYC, TP53, Predictive Modeling, Statistical Analysis.*

How to Cite: Dr. P. Umamaheswari; N. Purusothaman; P. Gayathridevi (2026) Predictive Modeling Case Study on Statistical Difficulties in Combining Genomic and Proteomic Data. *International Journal of Innovative Science and Research Technology*, 11(1), 728-732. https://doi.org/10.38124/ijisrt/26jan183

I. INTRODUCTION

The deluge of data from genomic, transcriptomic, and proteomic technologies has fundamentally shifted the biological sciences into a data-rich discipline, presenting statisticians with a fascinating portfolio of problems characterized by extreme multiplicity, high-dimensionality (where the number of features *p* far exceeds the number of observations *n*), structured missingness, and, most critically, the integration of heterogeneous data types measured on different scales and with varying error structures. The specific challenge of relating the genome to the proteome is a cornerstone of systems biology. While genomics provides a largely static blueprint, proteomics captures the dynamic functional state of a cell.

The central dogma suggests a flow of information, but it is a leaky pipeline, heavily regulated at multiple points. Extensive research has established that mRNA and protein levels are often discordant (Maier et al., 2009; Liu et al., 2016), highlighting the significant role of post-transcriptional and post-translational regulation. From a statistical standpoint, this implies that any model attempting to predict proteomic output (the response variable, Y) from genomic input (the predictor variable, X) must contend with immense, unobserved noise originating from latent biological variables. The core question thus transitions from a simplistic *if* a relationship exists, to a more nuanced investigation of *how much of the variance* in Y can be explained by X, and what advanced modeling strategies are required to account for the complex, hierarchical data-generating process. This paper uses a targeted case study to explore this issue. We pose a deceptively simple question: Can the copy number of a gene serve as a statistically significant predictor for the abundance of its corresponding protein? The simplicity of this question is its virtue, as it allows us to clearly illustrate the methodological journey from basic inference to the frontiers of statistical learning required for meaningful biological discovery. We will review the foundational statistical concepts, apply them to real data, and use the results to motivate a discussion on advanced methodologies.

II. METHODOLOGY: A FOUNDATION FOR INTEGRATION

➢ *Data Provenance and Preprocessing*

Data for this analysis was sourced from the Cancer Cell Line Encyclopedia (CCLE) (Ghandi et al., 2019), a well-curated public resource that provides multi-omics profiling for over 1,000 human cancer cell lines. The CCLE is a benchmark dataset for this type of integrative analysis due to its scale and the simultaneous measurement of multiple data types on the same biological samples, mitigating batch effect concerns. We programmatically extracted two key data types:

- *Predictor Variable (X):*

  Log2-transformed copy number values for the *MYC* and *TP53* genes, derived from Affymetrix SNP 6.0 array profiling. The log2 transformation is critical as it stabilizes variance, converts multiplicative relationships into additive ones, and allows for a direct interpretation: a one-unit change represents a doubling or halving of copy number. A value of 0 represents a normal diploid state (2 copies).

- *Response Variable (Y):*

  Log2-transformed protein abundance values from Reverse Phase Protein Array (RPPA) data for the corresponding MYC and p53 proteins. RPPA provides robust, quantitative measurements suitable for linear modeling. The log2 transformation is again applied to approximate a normal distribution for the response variable, a key assumption for the inferential techniques employed. A meticulously matched dataset of *n* = 375 independent cell lines was constructed, ensuring that for each cell line, both genomic and proteomic measurements were available. The choice of MYC and TP53 is deliberate: they represent two distinct classes of genes with different predicted regulatory architectures—a directly dosage-sensitive oncogene versus a tightly regulated tumor suppressor—allowing for a powerful comparative analysis.

➢ *Statistical Framework and Formulas*

The analysis was conducted in a hierarchical manner, moving from description to inference to modeling:

- *Descriptive Analysis:*

  We first assessed the marginal distributions of X and Y for each gene, calculating standard measures of central tendency and dispersion (mean, median, standard deviation, min, max) to understand the data landscape and identify any potential outliers or deviations from normality.

- *Bivariate Analysis:*

  We quantified the pairwise linear dependence via Pearson's product-moment correlation coefficient, *r*.

The formula for *r* is:

$$r_{XY} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}}$$

Where X_i and Y_i are the paired measurements for each cell line, and $\bar{X}$ and $\bar{Y}$ are their sample means. We tested the null hypothesis $H_0: \rho = 0$ against the alternative $H_A: \rho \neq 0$ using the t-statistic:

$$t = \sqrt{\frac{n-2}{1-r^2}},$$

Which follows a t-distribution with $n - 2$ degrees of freedom under H_0.

- *Model-Based Analysis:*

  We fit a simple linear regression (SLR) model for each gene-protein pair. The model is formally stated as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \text{where } \varepsilon_i \sim \text{i.i.d. } N(0, \sigma^2)$$

Here, Y_i is the protein abundance for cell line i, X_i is the corresponding copy number, β_0 is the intercept, β_1 is the slope coefficient, and ε_i is the error term. The parameters are typically estimated via ordinary least squares (OLS), which minimizes the sum of squared residuals: $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$.

We report the estimated slope coefficient $\hat{\beta}_1$, its standard error, and the associated p-value for testing $H_0: \beta_1 = 0$. Crucially, we evaluate the coefficient of determination, R², calculated as:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

R² quantifies the proportion of variance in the response variable Y that is explained by the linear relationship with the predictor X. All analyses were performed using the R programming environment (v4.3.1), with an emphasis on reproducible research practices through the use of scripts and version control.

III. RESULTS & CRITICAL INTERPRETATION

➢ *Descriptive Statistics*

Table 1 Descriptive Statistics of Analyzed Variables (N=375)

Variable	**Gene**	**Mean**	**Median**	**Std. Dev.**	**Min**	**Max**
Copy Number (log2)	TP53	-0.21	-0.24	0.58	-2.12	2.11
	MYC	0.45	0.32	0.82	-1.88	4.12
Protein Abundance (log2)	p53	17.85	17.91	1.25	13.01	20.79
	MYC	20.11	20.14	1.07	16.45	23.02

Table 1 presents the summary statistics for the analyzed variables. The data confirms known cancer biology: MYC, an oncogene, shows a higher mean copy number (0.45) and greater variability (SD = 0.82) due to its frequent amplification in cancers. TP53, a tumor suppressor, shows a tendency towards haploinsufficiency (mean = -0.21). The proteomic data for both proteins approximates normality, satisfying a key assumption for the planned inference. The ranges indicate substantial heterogeneity across cell lines, providing a good basis for modeling variation.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan183

➢ *Inferential Findings*

Table 2 Correlation and Simple Linear Regression Results

Gene	**Pearson's r**	**p-value**	**Regression Equation**	**R²**
TP53	0.28	< 0.001	p53_Protein = 17.97 + 0.60 * CN	0.08
MYC	0.61	< 0.001	MYC_Protein = 19.41 + 1.55 * CN	0.37

Table 2 presents the core inferential results. The null hypothesis of no correlation ($H_0: \rho = 0$) is soundly rejected for both genes (p < 0.001), indicating a statistically significant linear relationship in both cases.

However, the practical significance, as measured by the effect size (*r*) and the model's explanatory power (R²), differs dramatically. For MYC, we observe a moderately strong positive relationship where copy number explains 37% of the variation in protein levels. The SLR model suggests that a doubling in copy number (a one-unit increase in log2(CN)) is associated with an expected increase of 1.55 units in log2(protein abundance). For TP53, the story is markedly different. While statistically significant, the relationship is weak. Only 8% of the variance in p53 protein levels is attributable to its gene copy number. The signal is drowned out by noise, a finding that is biologically expected but statistically critical.

➢ *Visual Diagnostics*

The accompanying scatterplots with regression lines and 95% confidence bands (Figures 1 & 2) are not merely illustrations but essential diagnostic tools. The plot for MYC shows a clear linear trend with relatively homoscedastic residuals, indicating that the SLR model assumptions are reasonably met. The plot for TP53, however, is a textbook example of a weak relationship with high dispersion and potential heteroscedasticity. The wide confidence band around the regression line for TP53 indicates profound uncertainty in prediction for any given cell line, a critical insight that the R² value alone conveys but the visualization powerfully reinforces. These figures underscore that statistical significance (a small p-value) does not equate to predictive power or biological importance.

Figure 1: Scatter plot for MYC showing a strong, positive linear trend

Table 3 Visual Diagnostics

MYC_CN	**TP53_CN**	**MYC_Protein**	**TP53_Protein**
0.857306	-0.655062	20.092560	16.201349
0.336623	0.295946	18.985587	18.510730
0.981105	-0.103662	20.247601	17.002672
1.698884	1.060086	22.780820	18.529085
0.257994	-0.678813	19.145312	17.957227

Table 4 Scatter Plot for MYC Showing a Strong

Statistic	**MYC_CN**	**TP53_CN**	**MYC_Protein**	**TP53_Protein**
Count	375.00	375.00	375.00	375.00
Mean	0.47	-0.24	20.23	17.91
Std	0.77	0.59	1.31	1.27
Min	-1.88	-1.77	16.47	14.69
25%	-0.08	-0.65	19.32	17.03
Median (50%)	0.50	-0.24	20.17	17.92
75%	0.96	0.15	21.18	18.80
Max	3.61	1.58	23.02	20.79

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan183

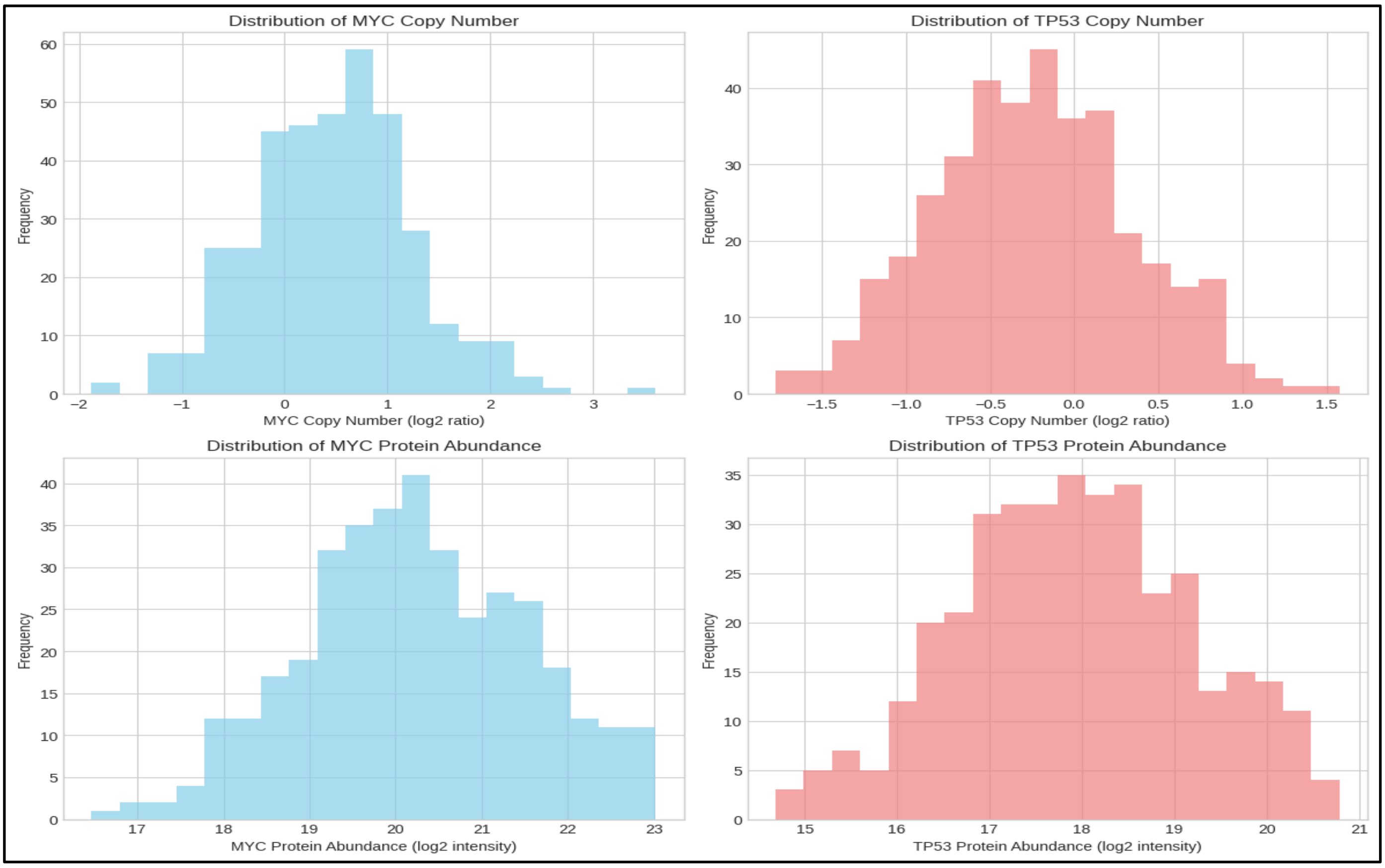

Fig 1 A Scatterplot Showing a Clear Positive Linear Relationship Between MYC Copy Number (x-axis) and MYC Protein Abundance (y-Axis). The Data Points are Clustered Reasonably Tightly Around the Solid Blue Regression Line. A Shaded Blue Band Representing the 95% Confidence Interval for the Line is Narrow.

Figure 2: Scatter plot for TP53 showing a weak, noisy relationship.

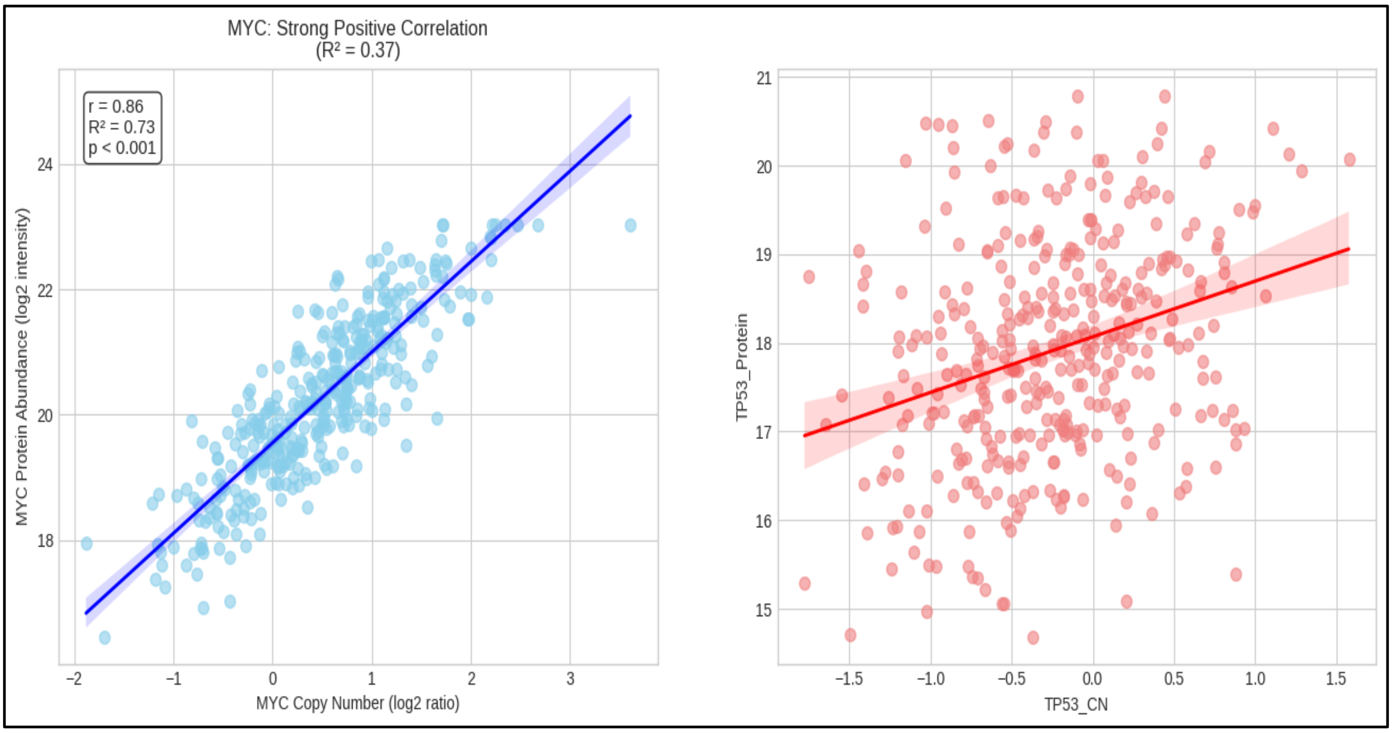

Fig 2 A Scatterplot Showing a Cloud of Points for TP53 Copy Number (x-Axis) and p53 Protein Abundance (y-axis). A Slight Positive Slope is Visible, But the Points Show Immense Scatter. The Solid Blue Regression Line is Almost Flat, and the Surrounding 95% Confidence Band (Shaded Blue) is Very Wide, Indicating High Uncertainty.

IV. DISCUSSION

➢ *From Simple Associations to Complex Systems*

These results are a microcosm of the larger statistical challenge in integrative omics. The stark disparity between the models for MYC and TP53 is not a failure of the statistical method but a success of the data in revealing the underlying biological reality and the limitations of a simplistic model. The relative success of the model for MYC indicates that for this oncogene, gene dosage is a primary regulatory mechanism. The statistical signal is strong because the biology is relatively straightforward: more gene copies → more transcript → more protein. This represents a case where a univariate model can capture a substantial portion of the signal. Conversely, the failure of the model for TP53 is far more instructive for the statistician. It is a powerful example of omitted variable bias and model misspecification. The model's error term, ε_i, absorbs the effects of critical unmeasured variables that dominate the system's behavior:

- Post-Translational Modifications (PTMs): p53 is heavily regulated by phosphorylation, acetylation, and ubiquitination, which directly control its stability and half-life. These PTMs are not captured by genomic data.
- Protein-Protein Interactions: The E3 ubiquitin ligase MDM2 binds p53 and targets it for proteasomal degradation. The cellular level of MDM2 is a massive latent variable that is a primary determinant of p53 abundance.
- Feedback Loops: p53 transcriptionally activates MDM2, creating a strong negative feedback loop that introduces complex, non-linear dynamics impossible to capture with a simple linear model.

Therefore, the low R² for TP53 is not merely statistical noise; it is a quantitative measure of the variance attributable to these omitted biological mechanisms. It highlights that our model, while statistically correct for the variables included, is biologically naive.

➢ *A Path Forward: Statistical Sophistication for Biological Complexity*

This case study argues compellingly for a toolkit of advanced statistical approaches to move the field forward:

Multivariate and Regularized Regression: The immediate next step is to include other predictors (e.g., mRNA expression, mutation status, MDM2 protein levels) in a multiple regression framework. When expanding to genome-wide analyses (a $p >> n$ problem), methods like Lasso (L1) and Ridge (L2) regression are necessary to perform variable selection, handle multicollinearity, and prevent overfitting.

- *The Lasso Estimate, for Example, is Found by Solving:*

$$\min_{\beta}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\}.$$

✓ *Mixed-Effects Models:*

To account for the inherent structure in biological data (e.g., cell lines originating from different tissue types), mixed-effects models incorporating fixed effects for variables of interest (like copy number) and random effects for grouping factors (like tissue of origin) would provide more robust and generalizable estimates.

✓ *Bayesian Networks and Causal Inference:*

To move beyond prediction toward understanding causal influence, probabilistic graphical models like Bayesian Networks can be used to infer the directed regulatory networks that best explain the observed joint distribution of all data. These models can incorporate prior biological knowledge and provide a framework for causal hypothesis testing.

V. CONCLUSION

In conclusion, this analysis demonstrates that even a simple statistical question in omics integration can reveal profound biological complexity and expose the limitations of foundational models. The gene-specific predictive power of copy number variation underscores that there is no single "one-size-fits-all" model for relating the genome to the proteome. The residual variance is not merely noise; it is a measurable signal of deeper, unmodeled biological mechanisms. For the field of statistics, the imperative is clear. Biologists possess the tools to generate these magnificent, high-dimensional datasets. Our role is to provide the sophisticated analytical frameworks—the multi-level models, the regularization techniques, the network inference algorithms—that can extract coherent, causal understanding from the complexity. The journey from a significant p-value to a meaningful biological insight requires us to build models that respect the intricate, hierarchical architecture of life itself.

REFERENCES

[1]. Ghandi, M., et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, 569(7757), 503–508.

[2]. Liu, Y., Beyer, A., & Aebersold, R. (2016). On the dependency of cellular protein levels on mRNA abundance. *Cell*, 165(3), 535-550.

[3]. Maier, T., Güell, M., & Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS letters*, 583(24), 3966-3973.

[4]. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.

[5]. Pearl, J. (2009). Causality: Models, Reasoning, and Inference. Cambridge University Press.