Discovery of Conditionally Independent Networks Among Gene Expressions in Breast Cancer Using Fast Step Graph

Grecia C. G. Rivera[1], **Juan G. Colonna**[1], **Marcelo Ruiz**[2]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Caixa Postal 69077-470 – Manaus – AM – Brasil

[2]Faculdade de Ciencias Exactas – Universidad Nacional de Río Cuarto (UNRC)
Ruta Nac. 36 - Km. 601, ZIP: X5804BYA - Córdoba - Argentina

{grecia.rivera; juancolonna}@icomp.ufam.edu.br
mruiz@exa.unrc.edu.ar

Abstract. *The heterogeneity of the causes of breast cancer and these complex gene interactions that characterize this neoplasm present significant challenges to understanding and treating the disease. This study is motivated by the need to identify interconnected networks of breast cancer genes, specifically those that represent conditional independence relationships. To construct these networks, we propose the use of the Fast Step Graph algorithm, which belongs to the family of sparse, high-dimensional Gaussian Graphical Models, applied to the PAM50 gene expression dataset. This dataset was stratified according to estrogen and progesterone receptors, key elements for prognosis and personalized therapy. The application of the algorithm resulted in four graphs highlighting the relationships among the genes involved in breast cancer. These findings support the hypothesis that there are specific gene sub-networks and contribute to a deeper understanding of gene interactions in this cancer, potentially offering new insights for future research and new therapeutic strategies.*

Resumo. *A heterogeneidade das causas do câncer de mama e as complexas interações gênicas que caracterizam essa neoplasia apresentam desafios significativos para a compreensão e o tratamento desta doença. Este estudo é motivado pela necessidade de identificar redes de genes interligados do cancer de mama, especificamente aquelas que representam relações de independência condicional. Para criar essas redes, propomos o uso do algoritmo Fast Step Graph, que pertence à família de Modelos Gráficos Gaussianos esparsos e de alta dimensionalidade, aplicado à base de dados de expressão gênica PAM50. Esta base de dados foi estratificada de acordo com os receptores de estrogênio e progesterona, elementos cruciais para o prognóstico e a terapia personalizada. A aplicação do algoritmo resultou na obtenção de quatro grafos que destacam as relações entre os genes envolvidos no câncer de mama. Esses achados apoiam a hipótese de que existem sub-redes gênicas específicas e contribuem para uma compreensão mais profunda das interações gênicas deste câncer, podendo oferecer novos insights para pesquisas futuras e novas estratégias terapêuticas.*

1. Introduction

Breast cancer is the leading cause of mortality among the female population in many regions of Brazil [Instituto Nacional de Câncer 2022]. Due to the heterogeneous nature of this disease, it exhibits a high degree of diversity among tumors, as well as among the individuals affected by cancer [Polyak 2011]. Because of this, molecular biology has developed various techniques to characterize the gene expression profile of these tumors, with the aim of classifying them, guiding treatment, and even predicting clinical outcomes [Valero and Álvarez 2013]. A successful technique is the PAM50 molecular test (*Prediction Analysis of Microarray*), which consists of a classifier that uses 50 specific gene expressions to categorize tumors into four subtypes: Luminal A, Luminal B, HER2+, and Basal-like [Parker et al. 2009, Liu et al. 2016, Okimoto et al. 2024].

One of the most challenging issues when dealing with a genomic database is the high dimensionality inherent to this type of dataset, where the dimension p of the vector $\mathbf{X} = (X_1, \ldots, X_p)$, whose entries represent the genetic profile of each patient, far exceeds the number of samples n, i.e. $p >> n$. It is in this scenario that Gaussian Graphical Models (GGM) become the most appropriate type of stochastic modeling to understand the conditional dependence relationships that may exist among the entries of the p-random vector $X_1, \ldots, X_p$, where each variable X_i represents the gene expression of the i-th gene and $\mathbf{X}$ has a multivariate Gaussian distribution [Lauritzen 1996, Koller and Friedman 2009, Liang and Jia 2023].

In GGMs, an edge of the graph that connects nodes i e j represents the existence of an association between X_i e X_j, conditional on the remaining variables. Therefore, the absence of an edge equates to the existence of conditional independence. Through appropriate parameterization, the set of conditionally independent variable pairs is represented by the number of zeros in the inverse—called the precision matrix—of the covariance matrix of the p-dimensional vector $\mathbf{X}$. A common assumption in high-dimensional statistics is that the graph is sparse or, equivalently, the precision matrix has most of its off-diagonal entries as zeros. Thus, sparse GGMs allow distinguishing significant (conditionally dependent) links from non-significant (conditionally independent) ones [Hastie et al. 2015]. It is also important to note that there is still no precise understanding of the conditional associations among the genes in the PAM50 breast cancer subset [Mendonca-Neto et al. 2022].

The limited understanding of gene associations, combined with the challenges of a high-dimensional setting ($p >> n$), motivates this study, which proposes the use of the Fast Step Graph algorithm−an optimized version of StepGraph− for efficient estimation of the sparse precision matrix (Ω) in PAM50 samples. The Fast Step Graph allows for the step-by-step identification of conditionally dependent gene pairs, offering a clear visualization of gene interactions through graphs [Zamar et al. 2021].

Therefore, this study represents one of many possible applications of the Fast Step Graph algorithm, demonstrating its potential in analyzing genomic data. Specifically, we applied it to real breast cancer data to determine the conditional dependence relationships among genes in patients with different hormonal biomarkers: estrogen receptor (ER) and progesterone receptor (PR) statuses. By identifying these relationships, we aim to deepen our understanding of breast cancer biology and reveal potential novel signaling pathways that could serve as innovative therapeutic targets.

2. Backgrounds

2.1. Gaussian Graphical Models in High-Dimensional Data

One of the most challenging issues when dealing with a genomic database is the high dimensionality inherent to this type of data, where the number of variables, corresponding to the number of genes, far exceeds the number of samples ($p >> n$), that is, the individuals or patients included in the study [Tsai et al. 2022]. The high dimensionality of these data does not allow traditional machine learning models to be fitted with confidence and without bias, and it also imposes substantial challenges concerning computational efficiency, given the considerable volume of relationships to be analyzed between pairs of variables.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_P)^T$ be a p-dimensional random vector following a Gaussian distribution with mean $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, assumed to be positive definite. The precision matrix is defined as $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Now, consider the data matrix $\mathbf{X}_{n\times p}$ where each row $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ corresponds to an independent observation of the random vector $\mathbf{X}$, with n representing the number of patients and p the number of genes.

The entries of the precision matrix Ω satisfy $w_{ij} = 0$ whenever variables X_i and X_j are conditionally independent given all other variables. Therefore, the matrix Ω can be used to represent the graph $G = (V, E)$ associated with the Gaussian Graphical Model (GGM), where only the edges corresponding to $w_{i,j} \neq 0$ are present. Here, the set of vertices is defined as $V = \{1, \ldots, p\}$ representing p genes, and the set of edges E consists of the pairs (i, j) were $i < j$ and X_i and X_j are conditionally dependent, given all the other variables variables [Dawid 1979, Lauritzen 1996, Yuan and Lin 2007, Liang and Jia 2023].

To estimate the matrix $\boldsymbol{\Omega}$ and thereby the graph $G = (V, E)$, we will employ the Fast Step Graph algorithm, which enables the estimation of all w_{ij} values, forming a graph that represents the relationships among the Gaussian variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$. In this context,the graph G represents a mathematical structure where the vertices $V = X_1, \ldots, X_p$ (or nodes), which in this case are genes, and edges E represent the conditional dependencies between ordered pairs (i, j) of nodes. If there is no edge between two nodes, which means $\Omega_{ij} = 0$, then these genes are conditionally independent given the rest.

In the context of breast cancer, the Fast Step Graph algorithm presents itself as a promising approach for identifying an optimal GGM that captures the associations underlying gene expression in cancerous cells. By employing this model, it is possible to identify genes with significant expression levels and discern those that exhibit conditional associations.

2.2. The Fast Step Graph Algorithm

The algorithm is a tool that enables the extraction of meaningful relationships in complex data by applying probabilistic modeling techniques, hyperparameter optimization, and structural learning.

Given this, the algorithm begins by considering an empty graph and enters an iterative cycle, passing through the *Forward* and *Backward* steps several times until it reaches a stopping condition. Figure 1 illustrates the evolution in the construction of the reference graph across different iterations indicated by the value of k. It is observed that

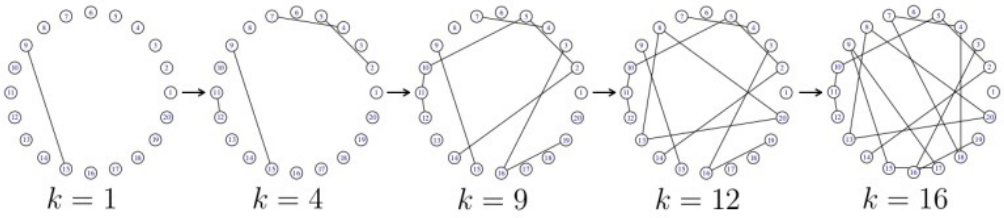

Figure 1. Graph construction using the *Step Graph*. k represents each iteration of the algorithm. Figure adapted from [Zamar et al. 2021].

in the final iteration, $k = 16$, the final graph is discovered. In this synthetic example, the "real graph" of the variables is known and serves to compare with the result obtained by the Fast Step Graph. However, when dealing with real data, this graph is unknown, and we can assume that the graph estimated by the procedure is a sufficiently good approximation. The Fast Step Graph algorithm, published on R-Cran[1], is based on the original *Step Graph* procedure and has been evaluated.

Forward Step: The *forward* step aims to create connections between variables. To do this, a linear regression is calculated between each variable indexed with a node and the remaining variables indexed with the nodes in its neighborhood. The residual error of this regression is used to calculate the correlations between each pair of variables. If the highest correlation identified between a pair of residuals exceeds a predefined threshold value called α_f, this connection is added to the graph of variables. This step is repeated iteratively until there are no more correlations whose values exceed the $\alpha_f$threshold.

Backward Step: This step is responsible for correcting the connections established during the previous (*forward*) phase.It checks if, after adding a new edge to the graph, the correlations of the previous residual errors, between the already connected nodes, have not fallen below a cutoff value called α_b. Those pairs of old connections whose correlations no longer meet this criterion are removed, while the connections that remain above α_b are retained in the graph.

Final Graph: The correlations used in the algorithm are calculated from the residuals resulting from the linear regressions between each variable and the variables in its neighborhood. Therefore, the resulting communities of nodes that remain in the final graph are formed solely by the variables that have strong conditional correlations, added in the *forward* step and not eliminated in the *backward* step.

2.3. Cross-Validation

The cross-validation (CV) method is used by the *CV Fast Step Graph* algorithm to find the values of the hyperparameters α_f and α_b, as these are not known a priori. This optimization is essential for obtaining the best estimate of Ω. During the search for the optimal α_f and α_b values, a grid of equally spaced values is used. The selection of the maximum and minimum of this grid is critical, as it can lead to a computationally expensive search or compromise the identification of the optimal values. In each iteration of the cross-validation, the algorithm assesses the performance of the GGM model by subdividing the data into *k-folds* (n-subsets) randomly, with some patients used for training and others for

[1] `https://cran.r-project.org/web/packages/FastStepGraph/index.html`

testing. In each step, $n-1$ sets are chosen for training, and the remaining set is reserved for testing.

The loss is calculated on the test set as the mean squared error between the model's predictions and the actual data of the variables (genes) in the test set. Therefore, the smaller the loss, the better the fit of the optimal α_f and α_b. Moreover, the threshold parameters α_f (forward) and α_b (backward) play a crucial role in determining the sparsity (or density) of the resulting graph. Lower values of α_f allow weaker conditional correlations to be included as edges, resulting in denser networks, while higher thresholds produce sparser graphs by including only the strongest associations. Similarly, α_b controls the removal of previously added edges; a lower α_b favors pruning and leads to sparser graphs. Therefore, the choice of these parameters has a direct effect on the topological structure of the estimated graph and must be carefully optimized.

3. Related Work

The fundamental idea behind sparse Gaussian graphical modeling is conditional independence, which allows for identifying only significant connections between variables. Motivated by this idea, [Yuan and Lin 2007] proposed a method for estimating the precision matrix Ω using $L1$ regularization. This penalization enforces sparsity by driving most coefficients in the precision matrix to zero.

[Friedman et al. 2008] proposed the *Graphical Lasso* (GLASSO) algorithm for constructing the precision matrix Ω, also using the $L1$ penalty, called *Lasso*. These authors built upon the "block coordinate descent" algorithm proposed by [Banerjee et al. 2008], who also sought to solve the maximum likelihood problem with $L1$ norm penalization. However, GLASSO has notable limitations, particularly its tendency to introduce false positives, leading to an overestimation of connections between variables.

Similarly, the CLIME algorithm proposed by [Cai et al. 2011] seeks to estimate precision matrices using the $L1$ minimization method for both sparse and non-sparse matrices. This algorithm guarantees the symmetry of the obtained precision matrix and decomposes the variables into subproblems, which are solved using linear programming. It is important to note that in an evaluation with numerical data, the Glasso method tends to include more non-zero elements compared to CLIME, which more closely resembles the true model. Additionally, in tests with real breast cancer data, the performance of CLIME is higher than the GLASSO method according to the Matthews Correlation Coefficient (MCC). The CLIME variant is even more computationally expensive than the original GLASSO version.

In contrast to these penalization-based methods, but with the same objective of identifying conditionally dependent variables, [Zamar et al. 2021] proposed Step Graph. This algorithm is based on the concept of Pearson correlation between the residuals of linear regressions and operates with a local search method that discovers the neighborhood of nodes in a sparse GGM model represented by its precision matrix. This method is based on the fundamental idea of an iterative algorithm that performs the *Forward* and *Backward* steps as described in the fundamentals section.

To compare the performance of these methods, [Maldonado and Ruiz 2022] evaluated GLASSO, CLIME, and Step Graph using a synthetic numerical dataset. Their re-

sults demonstrated that Step Graph outperforms both GLASSO and CLIME in terms of accuracy when estimating the precision matrix Ω. As part of their evaluation, they introduced a methodology to assess the performance of graph estimators, including the concept of non-informative estimators. Despite Step Graph's improved accuracy, its main drawback remains high computational cost, which can limit its applicability to large datasets.

One of the primary advantages of the Step Graph is its superior precision in reconstructing the true network structure by accurately selecting edges based on statistical relevance. While GLASSO tends to overestimate connections due to its L1 penalization—thereby introducing spurious edges and incorrect conditional dependencies—and CLIME, although producing sparse networks, may eliminate significant connections and incur high computational costs, Step Graph employs an iterative refinement process that effectively reduces both false positives and false negatives. This results in networks that are more interpretable and biologically meaningful, striking an optimal balance between sparsity and efficiency for large-scale applications, such as in the analysis of gene expression networks where the precise identification of conditional dependencies is crucial [Maldonado and Ruiz 2022].

Step Graph also exhibits greater adaptability across different graph structures. The performance of GLASSO is highly sensitive to graph density, struggling with both very dense and very sparse networks [Maldonado and Ruiz 2022]. In contrast, Step Graph provides more consistent performance across varying network topologies, making it particularly suitable for gene expression datasets such as PAM50, commonly used in breast cancer research. This flexibility ensures that Step Graph remains reliable even when applied to complex and heterogeneous biological data.

Despite the advantages of Step Graph in terms of accuracy and robustness, its high computational cost presents a challenge. Based on this review, we propose using the Fast Step Graph implementation, an algorithm that does not alter the original intention of the Step Graph algorithm but includes modifications in its implementation that substantially improve its computational performance, allowing us to tackle problems with more data, such as the case of the real PAM50 dataset.

4. Methodology

4.1. Database Description

The database used in this study was obtained from the Clinical Proteomic Tumor Analysis Consortium (CPTAC), which aims to accelerate the understanding of the molecular basis of cancer through large-scale analyses of the proteome and genome, known as proteogenomics [Mendonca-Neto et al. 2022, National Cancer Institute 2023]. From this consortium, the proteogenomic breast cancer (BRCA) database was chosen, which contains samples from 122 patients and 23,691 gene expressions (variables). These variables are pre-normalized, with a mean of zero and a standard deviation of one. PAM50 refers to the 50 most informative variables within the total set of 23,691 possible variables that have already been pre-selected for use in this work.

4.2. Data Preprocessing

The data preprocessing began with stratifying the patients using the positive and negative values of estrogen and progesterone biomarkers (ER/PR), which identify the presence

Table 1. Means of the hyperparameters α_f and α_b found with 5-folds Cross-Validation over 100 runs, with Confidence Intervals in parentheses.

Database (Strata)	Samples (Patients)	Mean of Alpha (CI)	Mean of Beta (CI)	Average Time (Seconds)
ER (+)	81	0.65 (0.00)	0.26 (0.02)	29.01 (4.79)
ER (-)	39	0.86 (0.00)	0.42 (0.01)	08.30 (1.46)
PR (+)	68	0.78 (0.00)	0.31 (0.02)	29.87 (4.77)
PR (-)	47	0.84 (0.01)	0.40 (0.01)	10.18 (1.77)

or absence of hormone receptors. After this stratification, four distinct gene expression matrices were obtained: (a) ER positive with 81 patients, (b) ER negative with 39 patients, (c) PR positive with 68 patients, and (d) PR negative with 47 patients, all containing 50 genes (or columns).

Each of these matrices was subjected to a standardization step, in which each gene (column) was normalized to have a mean of zero and a standard deviation of one using Z-score normalization (when the parameter `data_scale=TRUE`). This step prevents genes with large expression variances from disproportionately influencing the correlation-based edge selection in the algorithm. These standardized matrices were then used as input for the *Fast Step Graph* and *CV Fast Step Graph* algorithms to estimate genes that exhibit conditional dependencies and may be associated with the presence or absence of ER/PR receptors in the cancer cells of each patient.

4.3. Configuration and Execution of Experiments

The hardware used for the experiments was a laptop with an Intel Core i5 10210U processor, 8GB of RAM, a 256GB SSD, and the R programming language. The experiments are available on GitHub (https://github.com/gregscristina/faststepgraph-application).

The first method executed was the CV Fast Step Graph with the aim of finding the optimal values for the hyperparameter α_f and α_b. To ensure a fair comparison and avoid potential biases from data sampling strategies, the same data division was consistently applied across all tables during hyperparameter evaluation. Specifically, a 5-fold cross-validation procedure was performed 100 times with different random seeds for each stratified dataset, ensuring convergence to the average optimal values. The hyperparameter grid interval was $0,5 \leq \alpha_f \leq 0,9$ and $0,25 \leq \alpha_b \leq 0,45$. The results are shown in Table 1 which presents the mean optimal values along with their confidence intervals (CI) at a 0.05 significance level. The last column of the table provides the average execution time per iteration of the algorithm.

After calculating the optimal parameters, these were used to execute the Fast Step Graph to obtain the four sparse precision matrices: Ω_{ER+}, Ω_{ER-}, Ω_{PR+} and Ω_{PR-}. The resulting graphs, representing the inferred conditional dependence structures, are illustrated in the next section.

5. Results

From the execution of the Fast Step Graph algorithm, we obtained four graphs, where each gene is represented by a node with its respective name, and each edge represents the

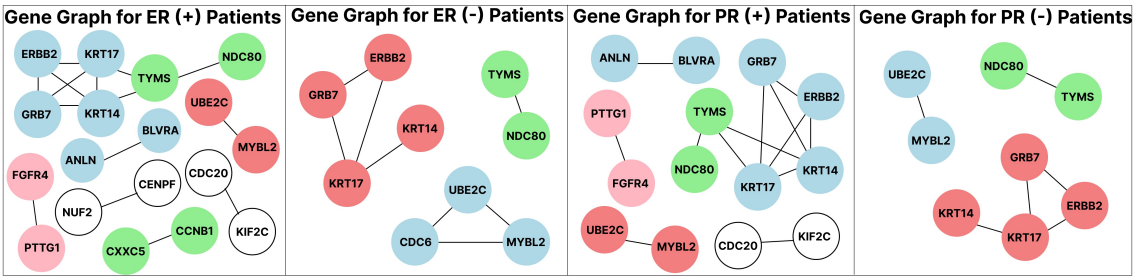

Figure 2. Gene interaction graphs by ER/PR strata in PAM50 breast cancer dataset.

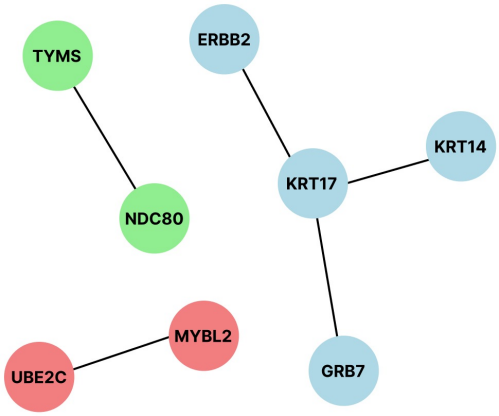

Figure 3. Consistently identified gene communities across all four PAM50 strata.

existence of a estimated conditional dependence between pairs of genes. Figure 2 illustrates the relationship graphs between genes on each strata of the PAM50 dataset–ER(+), ER(-), PR(+) and PR(-). The pairs of genes represented are those that the algorithm identified as significant and conditionally dependent, which, in turn, manifest in patients with or without estrogen receptors (positive or negative).

In these graphs, the communities formed by the most prominent genes can be observed. The most notable community is formed by the genes: NDC80, TYMS, KRT14, KRT17, GRB7, and ERBB2, which is present in patients with ER+ and PR+ hormone receptors, indicating a stronger conditional correlation among these cancer-related genes. It is important to note that each of these biomarkers regulates important biological processes and has scientific evidence related to breast cancer, which will be discussed in the next section.

From the results obtained, it is notable that certain gene communities are repeated in all four graphs. In Figure 3, the common conditional correlations across the four strata of the PAM50 dataset were extracted. For example, the relationship between the TYMS and NDC80 genes appeared in both positive and negative ER/PR patients. However, in positive ER/PR patients, this connection is part of a larger gene community, while in negative ER/PR patients, this connection appears in isolation.

UBE2C and MYBL2 Community: Based on the analysis of the four resulting graphs, we identified the common association of the MYBL2 gene with UBE2C. The overexpression of the MYBL2 gene is generally associated with biological processes

such as intensified cellular replication and genomic instability, hallmark characteristics of various types of cancer, including breast cancer [Razera et al. 2023]. The literature indicates that MYBL2 expression is directly related to unfavorable prognoses of breast cancer. Taking into account research on other neoplasms, it is observed that, in the case of gastric cancer, MYBL2 establishes a direct link with the promoter region of UBE2C, activating its transcription [Long et al. 2023]. This process results in inhibition of apoptosis, conferring resistance to Cisplatin, a chemotherapeutic agent that interferes with cell multiplication. It should be noted that the interaction between MYBL2 and UBE2C is not unidirectional, as UBE2C can also regulate the expression of MYBL2. Furthermore, the relationship between MYBL2 and UBE2C is also observed in a variety of cancers [Dastsooz et al. 2019]. Therefore, these consistencies suggest that the relationship found by our algorithm is relevant and indicate that the proposed methodology could be extended to other types of cancer.

TYMS and NDC80 Community: According to the literature, elevated levels of TYMS expression are associated with TNBC (triple negative breast cancer). Furthermore, TYMS is positively correlated with more aggressive characteristics, larger tumors, and negativity for estrogen and progesterone receptors [Song et al. 2021]. The presence of TYMS is not related to a specific type of breast cancer, which is why it is present on all four graphs, as a possible indicator of prognoses of breast neoplasia [Song et al. 2021].

The literature indicates that the NDC80 gene could be a possible target for TNBC treatment due to its relation to the mechanism of Cisplatin resistance [Li et al. 2022]. Other studies have shown that the NDC80 gene is related to the transition from normal tissues to benign tumors and that this same gene is related to multigenic expression profiles, often used in clinical settings to characterize breast cancer tissues and personalize therapies [Bièche et al. 2011, Koleck and Conley 2016].

Although no direct link between TYMS and NDC80 has been identified in the literature, the presence of both in all graphs suggests the possibility of indirect interactions or independent contributions to breast neoplasia. The complex and multifactorial nature of cancer often challenges the immediate identification of all gene interrelations. Thus, individual findings regarding the association of TYMS and NDC80 with breast cancer deserve attention for future research, especially in a context where their presence is consistent across different data sets.

GRB7, ERBB2, KRT17, and KRT14 Community: The literature establishes an association between the ERBB2 gene and the prognosis of breast cancer [Valero and Álvarez 2013]. Its overexpression is recognized for contributing to the uncontrolled growth of cancer cells, making it an important target in therapies. Additionally, this gene is also known for its association with recurrent alterations in brain metastases, occurring in 35% of cases [Priedigkeit et al. 2017]. The GRB7 gene is also associated with alterations in brain metastases, suggesting greater tumor aggressiveness. In contrast, the KRT17 and KRT14 genes show a decrease in cases of brain metastases. It is presumed that the identification of this gene community, found using our method, could raise new hypotheses for developing specific therapies, aiming to better understand from a molecular standpoint why these interactions manifest in breast cancer.

6. Conclusions

This study presented the Fast Step Graph algorithm not only as a methodological contribution for sparse precision matrix estimation in high-dimensional Gaussian graphical models, but also as a practical tool for discovering biologically meaningful structures in genomic data. By applying FSG to the PAM50 breast cancer dataset, stratified by ER/PR status, the algorithm revealed gene communities that consistently appeared across multiple runs and align with findings in the biomedical literature. Notably, connections such as **MYBL2–UBE2C**, **TYMS–NDC80**, and **GRB7–ERBB2–KRT17–KRT14** reflect known mechanisms of tumor progression, prognosis, and therapy resistance. These findings underscore both the potential of the Fast Step Graph algorithm in uncovering established interactions and its capacity to highlight previously undocumented relationships, thereby suggesting new directions for breast cancer research.

Designed to efficiently estimate sparse precision matrices in high-dimensional Gaussian settings, FSG proved to be a computationally effective and biologically meaningful approach to uncover conditional dependencies in genomic data. Its promising performance in this context demonstrates its potential for broader applicability across various domains where the discovery of conditional independence structures is critical.

Future research will include systematic comparisons with other graphical modeling methods, such as Weighted Gene Co-expression Network Analysis (WGCNA) [Langfelder and Horvath 2008], which constructs gene networks based on pairwise correlations. Contrasting the conditional dependency structures revealed by Fast Step Graph with the modules derived from WGCNA could provide deeper insights into the architecture of gene regulation and reveal complementary patterns not captured by a single approach. Additionally, future work will explore the performance of the Fast Step Graph algorithm on different types of genomic datasets, including RNA-seq and single-cell data, and continue refining its parameter tuning strategies. By enabling the discovery of both well-established and novel relationships, the Fast Step Graph algorithm opens up new possibilities and understanding of complex high-dimensional data in biomedical applications and beyond.

7. Acknowledgements

The present work is the result of the Research and Development (R&D) project 001/2020, signed with the Federal University of Amazonas and FAEPI, Brazil, which has funding from Samsung, using resources from the Informatics Law for the Western Amazon (Federal Law nº 8.387/1991), and its dissemination is in accordance with article 39 of Decree No. 10.521/2020. This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES-PROEX) – Funding Code 001. Additionally, this work was partially funded by the Foundation for Research Support of the State of Amazonas – FAPEAM – through the PDPG project. We would like to extend our thanks to Pedro Henrique Cassiano Cipriano for his valuable assistance regarding the considerations on genetics.

References

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516.

Bièche, I., Vacher, S., Lallemand, F., Tozlu-Kara, S., Bennani, H., Beuzelin, M., Driouch, K., Rouleau, E., Lerebours, F., Ripoche, H., Clairac, G., Spyratos, F., and Lidereau, R. (2011). Expression analysis of mitotic spindle checkpoint genes in breast carcinoma: Role of ndc80/hec1 in early breast tumorigenicity, and a two-gene signature for aneuploidy. *Molecular cancer*, 10:23.

Cai, T., Liu, W., and Luo, X. (2011). A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Dastsooz, H., Cereda, M., Donna, D., and Oliviero, S. (2019). A comprehensive bioinformatics analysis of ube2c in cancers. *International Journal of Molecular Sciences*, 20:2228.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(1):1–15.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Instituto Nacional de Câncer (2022). Dados e números sobre câncer de mama - relatório anual 2022. Relatório anual, Coordenação de Prevenção e Vigilância, Divisão de Detecção Precoce e Apoio à Organização de Rede, Rio de Janeiro, Brasil. Acesse: `www.inca.gov.br/mama`.

Koleck, T. and Conley, Y. (2016). Identification and prioritization of candidate genes for symptom variability in breast cancer survivors based on disease characteristics at the cellular level. *Breast Cancer: Targets and Therapy*, 8:29.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT press.

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Li, J., Xu, X., and Peng, X. (2022). Ndc80 enhances cisplatin-resistance in triple-negative breast cancer. *Archives of Medical Research*, 53(4):378–387.

Liang, F. and Jia, B. (2023). *Sparse Graphical Modeling for High Dimensional Data: A Paradigm of Conditional Independence Tests*. CRC Press.

Liu, M. C., Pitcher, B. N., Mardis, E. R., Davies, S. R., Friedman, P. N., Snider, J. E., Vickery, T. L., Reed, J. P., DeSchryver, K., Singh, B., et al. (2016). Pam50 gene signatures and breast cancer prognosis with adjuvant anthracycline-and taxane-based chemotherapy: correlative analysis of c9741 (alliance). *NPJ breast cancer*, 2(1):1–8.

Long, J., Zhu, B., Tian, T., Ren, L., Tao, Y., Zhu, H., Li, D., and Xu, Y. (2023). Activation of ubec2 by transcription factor mybl2 affects dna damage and promotes gastric cancer progression and cisplatin resistance. *Open Medicine*, 18(1).

Maldonado, J. and Ruiz, S. (2022). Assessment of covariance selection methods in high-dimensional gaussian graphical models. *Trends in Computational and Applied Mathematics*, 23:583–593.

Mendonca-Neto, R., Reis, J., Okimoto, L., Fenyö, D., Silva, C., Nakamura, F., and Nakamura, E. (2022). Classification of breast cancer subtypes: A study based on representative genes. *Journal of the Brazilian Computer Society*, 28(1):59–68.

National Cancer Institute (2023). Clinical proteomic tumor analysis consortium (cptac). Technical report, U.S. Department of Health and Human Services. Acesse: `https://gdc.cancer.gov/about-gdc/contributed-genomic-data-cancer-research/clinical-proteomic-tumor-analysis-consortium-cptac`.

Okimoto, L. Y. S., Mendonca-Neto, R., Nakamura, F. G., Nakamura, E. F., Fenyö, D., and Silva, C. T. (2024). Few-shot genes selection: subset of pam50 genes for breast cancer subtypes classifcation. *BMC Bioinformatics*, 25:92.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160.

Polyak, K. (2011). Heterogeneity in breast cancer. *The Journal of Clinical Investigation*, 121(10):3786–3788.

Priedigkeit, N., Hartmaier, R. J., Chen, Y., Vareslija, D., Basudan, A., Watters, R. J., Thomas, R., Leone, J. P., Lucas, P. C., Bhargava, R., Hamilton, R. L., Chmielecki, J., Puhalla, S. L., Davidson, N. E., Oesterreich, S., Brufsky, A. M., Young, L., and Lee, A. V. (2017). Intrinsic subtype switching and acquired erbb2/her2 amplifications and mutations in breast cancer brain metastases. *JAMA Oncology*, 3(5):666–671.

Razera, A., Rodrigo Santos, J., Gonçalves, L., Marques, L., Chao, B., Campos, D., and Carraro, E. (2023). Mybl2 gene as prognostic biomarker in breast cancer: A systematic review. *Journal of Advances in Medicine and Medical Research*, 35:101–107.

Song, S., Tian, B., Zhang, M., Gao, X., Jie, L., Liu, P., and Li, J. (2021). Diagnostic and prognostic value of thymidylate synthase expression in breast cancer. *Clinical and Experimental Pharmacology and Physiology*, 48(2):279–287.

Tsai, K., Koyejo, O., and Kolar, M. (2022). Joint gaussian graphical model estimation: A survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 14(6):e1582.

Valero, V. and Álvarez, R. H. (2013). Biologia molecular do câncer de mama. In *Tratado de Oncologia*, pages 2027–2052. Atheneu.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.

Zamar, R., Ruiz, M., Lafit, G., and Nogales, J. (2021). A stepwise approach for high-dimensional gaussian graphical models. *Journal of Data Science, Statistics, and Visualisation*, 1(2).