Unravelling breast and prostate common gene signatures by Bayesian network learning*

João Villa-Brito[1], Marta B. Lopes[2][0000−0002−4135−1857], Alexandra M. Carvalho[3][0000−0001−6607−7711], and Susana Vinga[2,4][0000−0002−1954−5487]

[1] Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
[2] IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3] Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa Av. Rovisco Pais, 1049-001 Lisboa, Portugal
[4] INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, R. Alves Redol 9, 1000-029 Lisboa, Portugal.
{susanavinga,alexandra.carvalho}@tecnico.ulisboa.pt

Abstract. Breast invasive carcinoma (BRCA) and prostate adenocarcinoma (PRAD) are two of the most common types of cancer in women and men, respectively. As hormone-dependent tumours, BRCA and PRAD share considerable underlying biological similarities worth being exploited. The disclosure of gene networks regulating both types of cancers would potentially allow the development of common therapies, greatly contributing to disease management and health economics. A methodology based on Bayesian network learning is proposed to unravel breast and prostate common gene signatures. BRCA and PRAD RNA-Seq data from The Cancer Genome Atlas (TCGA) measured over $\sim$ 20000 genes were used. A prior dimensionality reduction step based on sparse logistic regression with elastic net penalisation was employed to select a set of relevant genes and provide more interpretable results. The Bayesian networks obtained were validated against information from STRING, a database containing known gene interactions.

Keywords: sparse logistic regression · gene expression · machine learning.

1 Scientific Background

Due to the computerisation of our everyday life, available data is growing tremendously across all fields of research, businesses and industry. When dealing with

* Supported by the EU Horizon 2020 research and innovation program (grant No. 633974 - SOUND project), and the Portuguese Foundation for Science & Technology (FCT), through projects UID/EMS/50022/2019 (IDMEC, LAETA), UID/EEA/50008/2019 (IT), UID/CEC/50021/2019 (INESC-ID), PERSEIDS (PTDC/EMS-SIS/0642/2014), PREDICT (PTDC/CCI-CIF/29877/2017), NEUROCLINOMICS2 (PTDC/EEI-SII/1937/2014), and IF/00653/2012; also partially supported by internal IT projects QBigData and RAPID.

high-dimensional data, sparse models are able to extract knowledge from data, by identifying a smaller number of relevant variables (from a whole set of variables) explaining the data. In the context of biological data, the use of sparse graphical models is expected to disclose valuable insights on the underlying biological mechanisms.

This work searches for common gene signatures between breast invasive carcinoma (BRCA) and prostate adenocarcinoma (PRAD), two of the most common types of invasive cancer in women and men, respectively. Although arising in organs with different anatomies and physiological functions, BRCA and PRAD tumours depend on gonadal steroids for their development, as the organs they originate from, being hormone-dependent. Both cancers have considerable underlying biological similarities worth being exploited with the goal of improving patient outcomes [1]. The proposed methodology uses Bayesian network learning to identify a common gene network to both cancers, as not only the genes regulating the diseases but also the interaction between them could help better understanding the diseases, while providing guidance to cancer therapy research and disease management.

2 Materials and Methods

2.1 Dimensionality Reduction

Let Y be a random variable whose n components are independently distributed with means μ, X the $n \times p$ matrix containing the set of p explanatory variables, and $\beta = \{\beta_1, \ldots, \beta_p\}^T$ the $p \times 1$ vector of unknown regression coefficients associated with each covariate. Then, in a generalised linear model (GLM) [2]:

$$E(Y) = \eta = g(\mu) = X\beta = x_i\beta; \qquad i = 1, \ldots, n. \tag{1}$$

Specifically in this work, the independent variable is binary, thus Y is assumed to follow a Binomial distribution. Defining the probability of success, p_i as the probability of $Y_i = 1$, given the associated variables vector x_i, binary logistic regression (LR) models how the response variable depends on the set of variables:

$$\eta = logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = x_i\beta; \qquad i = 1, \ldots, n. \tag{2}$$

The unknown regression coefficients β are estimated using maximum likelihood. For a n sized sample, the log-likelihood function for a binary LR is

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]. \tag{3}$$

The estimates $\hat{\beta}$ obtained maximise $\ell(\beta)$. Usually, they are all non-zero, and if $p > N$ (more explanatory variables than observations), they are not unique.

When addressing healthcare big data problems, it is necessary to constrain the regression problem in order to estimate interpretable models, e.g. through regularised optimisation. In other words, it is necessary to encourage sparsity. In a sparse statistical model, only a relatively small number of predictors is different from zero. The more sparse the model is, the less the number of non-zero parameters.

Ridge regularisation [3] adds an ℓ_2 constraint, $\sum_{j=1}^{p} \beta_j^2$, to the log-likelihood function, promoting solutions with small norms, i.e., close to zero, but still non-sparse. The least absolute shrinkage and selection operator (lasso) [4] is a regularisation method that enjoys the stability of ridge regression, while promoting variable selection. It works by combining the log-likelihood function with an ℓ_1 constraint, $\sum_{j=1}^{p} |\beta_j|$.

The lasso penalty is able to perform both shrinkage and variable selection. While it performs well in many circumstances, it has shown some limitations. If $p > n$ the lasso selects no more than n variables before it saturates; if there are highly correlated variables, the lasso arbitrarily selects only one, not taking into account the group as a whole (there is no clustering); and when the variables are highly correlated, in $n > p$ situations, the prediction accuracy of the lasso becomes dominated by the ridge regression.

Elastic net regularisation [5] is a technique proposed to solve the mentioned problems, as a combination of both lasso and ridge, and performing as well as the lasso whenever the lasso does the best. The regulariser is defined as $\lambda \sum_{j=1}^{p} \left\{(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right\}$. The ℓ_1 part of the penalty helps to generate a sparse model, while the ℓ_2 part makes it possible to select more than n variables (in the $p > n$ case) and encourages clustering. The tuning constants, $\lambda \geq 0$ and $\alpha \in [0, 1]$ control the magnitude of the parameters and the relative weight of each constraint, respectively.

2.2 Bayesian Networks

Graphical models are a powerful probabilistic representation that provide interpretable models of the domain. For this reason, they have been used in a large variety applications such as genetics, oncology, computational biology, and medicine and health care. The large volume of high-dimensional biological data has motivated the use of graphical models to provide understanding into novel biological mechanisms.

Bayesian networks are the most widely known directed graphical models, however, they are typically not used with high dimensional data, with $p \gg n$, as they do not scale well with the number of variables. Nonetheless, these directed models provide us with unprecedented insights about probabilistic correlations between variables under study, in this case, gene expression values. This could be of great benefit for genomics applications, with datasets such as the human transcriptome , with $p \sim 20000$.

Considering a p-dimensional random vector $\mathbf{Z} = (Z_1, \ldots, Z_p)$, whose realisation is in X, a Bayesian network (BN) is rigorously defined as a directed acyclic graph (DAG) $G = (V; E)$ with nodes in V, coinciding with $\mathbf{Z}$, and edges in E,

representing a joint probability distribution $P(\mathbf{Z})$ in a factored way, according to the DAG structure as:

$$P(Z_1, \ldots, Z_p) = \prod_{j=1}^{p} P(Z_j | pa(Z_j), \theta_j), \tag{4}$$

where $\text{pa}(Z_j) = \{Z_i : Z_i \to Z_j \in E\}$ is the parent set of Z_j and θ_j encodes the parameters that define the conditional probability distribution (CPD) for Z_j. Gaussian CPDs for continuous data are considered.

Learning a BN reduces to learn its structure and parameters. Having the structure fixed, parameters are quite easy to learn. The hard task is to learn the structure itself, generally approached through score-based learning; in this case, a score is used to ascribe the network fitting to the data. Most common scoring criteria are based on maximum likelihood estimation with penalisation factors to prevent data overfitting.

Aragam et al. (2017) developed a new R package, called `sparsebn` [6], focused on learning the structure of sparse graphical models, especially thought for large networks. To learn a BN from data, they have used a score-based approach that relies on regularised maximum likelihood estimation. The following criterion was considered:

$$\min_{B \in D} \ell(B; X) + \rho_\lambda(B), \tag{5}$$

where ℓ denotes the negative log-likelihood, ρ_λ is some regulariser, the matrix B is the weighted adjacency matrix of a DAG, being D the set of weighted adjacency matrices that represent DAGs. For continuous data, a Gaussian likelihood with ℓ_1 or minimax concave penalty is used.

The package offers methods to learn the structure of a BN, to estimate its parameters $\hat{B}$, to plot that structure and, for Gaussian data, to calculate the implied covariance and precision matrices. Many methods from the literature on coordinate descent such as *warm starts*, *active set iterations*, *block updates* and *sparse data structures* were used by the authors to make the algorithms run faster, distinguishing `sparsebn` from existing packages, for sparse structure learning and high dimensional data.

2.3 Datasets

To unravel common gene signatures to breast and prostate cancers, two datasets were extracted from the Cancer Genome Atlas (TCGA) database [7]: BRCA and PRAD datasets, corresponding to breast and prostate, respectively. For more information on the datasets refer to https://github.com/jvillabrito/common-gene-signature.

A subset of 19810 variables was selected from the BRCA dataset, and of 19660 from the PRAD dataset, corresponding to the protein-coding genes reported from the Ensembl genome browser [8] and the Consensus CDS [9] project. The data was pre-processed as follows (Fig. 1). The variables with zero standard deviation were excluded from both datasets, and only the 19529 common to both

datasets were considered for further analysis. The variables were log-transformed and normalised to zero mean and unit variance. The final datasets are $X_{brca} \in R^{n_{brca} \times p}$; $y_{brca} \in R^{n_{brca}}$ and $X_{prad} \in R^{n_{prad} \times p}$; $y_{prad} \in R^{n_{prad}}$, with $n_{brca} = 1204$, $n_{prad} = 547$ and $p = 19529$. Matrices X are the explanatory variables (genes) matrices and vectors y are the binary response vectors, with '1' and '0' corresponding to tumour and normal tissue samples, respectively. Samples presented with metastases were not considered for the analysis.

2.4 Finding Common Gene Signatures

With the goal of obtaining more interpretable results, a dimensionality reduction step was added before learning the Bayesian networks, using logistic regression with elastic net penalisation, considering two values of α ($\alpha = 0.1$ and $\alpha = 0.01$). Two approaches were tested: *jointEN* and *sepEN*. In the first, sparse logistic regression with elastic net penalty was applied to a new dataset combining BRCA and PRAD data, BRCAPRAD ($X_{brcaprad} \in R^{n_{brcaprad} \times p}$; $y_{brcaprad} \in R^{n_{brcaprad}}$, with $n_{brcaprad} = 1751$ and $p = 19529$), using the `glmnet` R package. The λ parameter was optimised by 10-fold cross-validation. The variables selected were used for further analysis. In the second approach, two sparse logistic models were fit, one for BRCA and another for PRAD, also with 10-fold cross-validation. The variables used are the ones selected separately for each cancer.

After the dimensionality reduction block (Fig. 1), `sparsebn` R package was used to learn the Bayesian networks, using the method `estimate.dag`. The parameter 'edge.threshold' was used to force the number of edges in the solution to be less or equal than the number of nodes. The output of the method is a solution path, rather than an unique solution, consisting of a sequence of estimates for a predetermined set of lambdas $\lambda_{max} > \lambda_1 > \ldots > \lambda_{min}$ (default grid of values are used based on a decreasing log-scale). As λ decreases, there is less regularisation, i.e. the graphs are more dense, containing more edges. The `select.parameter` method was then used to get the optimal value of λ, based on a trade-off between the increase in log-likelihood and the increase in complexity between solutions. Only the solution for the optimal lambda was considered for further analysis. Four Bayesian networks were learnt: two from BRCA data, one using only tumour tissue samples (*tumour*BN) and another using only normal tissue samples (*normal*BN), and the same from PRAD data. The four Bayesian Networks obtained were validated by comparing the resulting edges with STRING information [10]. The *tumour*BNs were then compared to determine the number of shared edges. The same was done for *normal*BNs. Finally, *tumour*BNs were compared against *normal*BNs, to verify whether they share common edges.

3 Results

The number of edges of the Bayesian networks obtained can be found in Tab. 1, for reduced data and for full dimension data. In the case of reduced data, *jointEN*

and *sepEN* approaches are discriminated. BN_{brca} and BN_{prad} correspond to the Bayesian networks learnt from BRCA and PRAD data, respectively. It can be noticed that the solutions for the optimal lambdas have the number of edges close to the number of variables. A considerable overlap between BN_{brca} and BN_{prad} networks (# common edges; Tab. 1) was obtained, which supports the fact that both types of cancer have underlying similarities. A noticeable overlap was also obtained when comparing the pairwise gene connections identified and STRING information. The percentage of known gene interactions found when learning the BNs from tumour data is approximately twice the number of gene interactions when BNs are learnt from normal data.

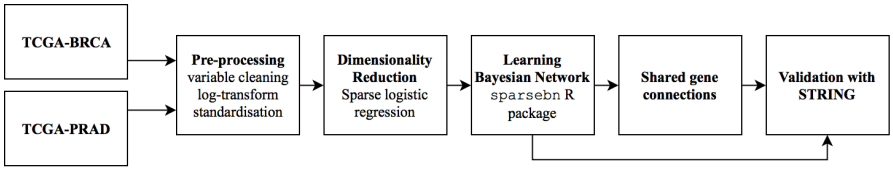
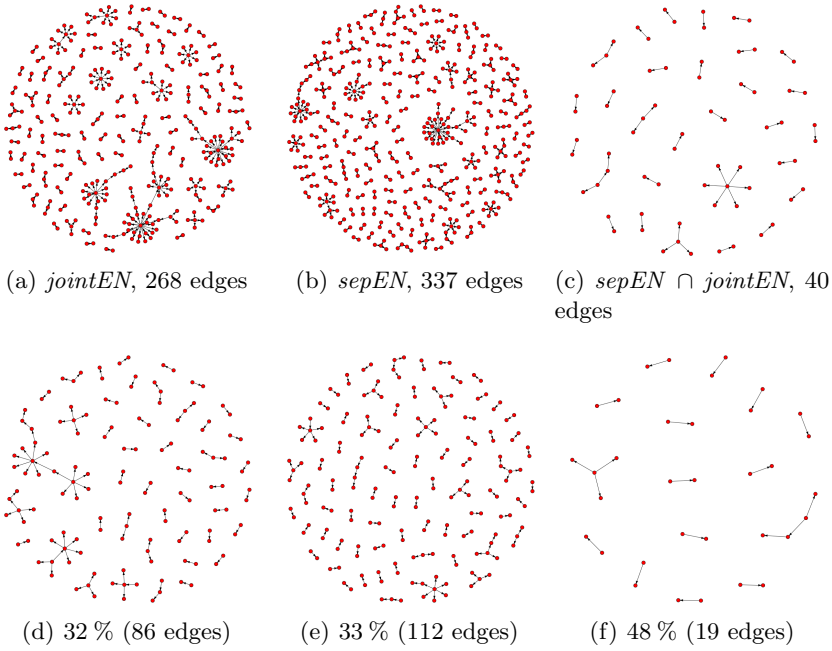


Fig. 1. Proposed solution pipeline.

Table 1. Number of edges of BNs obtained. The numbers in parentheses state the percentage of edges that represent known gene interactions, based on STRING information. (T: Tumour; NT: Non-Tumour)

α	data	n brca	n prad	approach	p	# edges BN_{brca}	# edges BN_{prad}	# common edges
0.1	T	1091	495	*jointEN*	546	499 (12 %)	528 (11 %)	57 (47 %)
				sepEN	738	624 (11 %)	733 (11 %)	33 (45 %)
	NT	113	52	*jointEN*	546	537 (7 %)	534 (8 %)	38 (21 %)
				sepEN	738	713 (9 %)	721 (7 %)	33 (24 %)
0.01	T	1091	495	*jointEN*	2791	2674 (16 %)	2663 (13 %)	268 (32 %)
				sepEN	4370	4125 (17 %)	4148 (14 %)	337 (33 %)
	NT	113	52	*jointEN*	2791	2790 (6 %)	2791 (9 %)	153 (18 %)
				sepEN	4370	4370 (8 %)	4336 (9 %)	287 (16 %)
–	T	1091	495	–	19529	18411 (23 %)	18553 (17 %)	1589 (31 %)
	NT	113	52	–	19529	19527 (9 %)	19366 (10 %)	1664 (16 %)

Figure 2 illustrates the networks of the common edges in BN_{brca} and BN_{prad}, when BNs are learnt from tumour data, after dimensionality reduction with $\alpha = 0.01$. Besides paired genes, highly connected genes were obtained as well, called hubs, which are also reported in STRING.



(a) *jointEN*, 268 edges (b) *sepEN*, 337 edges (c) *sepEN* ∩ *jointEN*, 40 edges

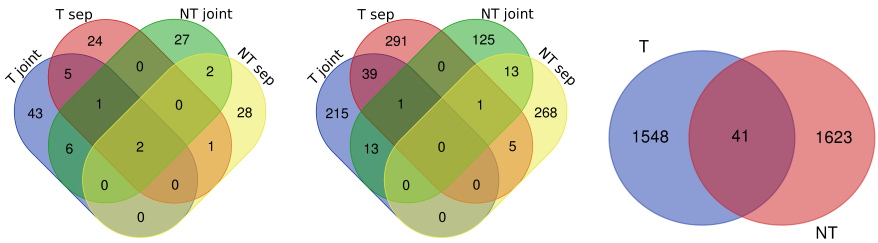
(d) 32 % (86 edges) (e) 33 % (112 edges) (f) 48 % (19 edges)

Fig. 2. Networks of common genes in BN_{brca} and BN_{prad} learnt form tumour data, by (a) *jointEN* and (b) *sepEN*, for $\alpha = 0.01$; (c) is the intersection of edges in (a) and (b); (d), (e), and (f) correspond to the edges from the networks in (a), (b), and (c) that are reported in STRING, respectively.

To infer whether the *tumour*BNs obtained are specific to BRCA and PRAD diseases or not, Venn diagrams were produced to illustrate the overlap between *tumour*BNs and *normal*BNs (Fig. 3). For $\alpha = 0.1$, the overlap is of 9 and 3 edges for *jointEN* and *sepEN*, respectively, while for $\alpha = 0.01$ the overlap is of 14 and 6 edges. With no regularisation, 41 edges in the *tumour*BN were also found in *normal*BN. These edges are more likely related to cell machinery, and therefore of little interest, and not related to BRCA and PRAD diseases. Such little overlap might be an indicator of the specificity of the *tumor*BNs obtained to the diseases under study.

4 Conclusion

The methodology proposed was able to extract common gene signatures to both types of cancer, BRCA and PRAD, by Bayesian network learning. A considerable overlap between the gene networks identified and STRING network information was obtained, a strong indication that the networks learnt may be biologically meaningful. The present results are expected to play a role in cancer therapy research, by fostering cancer therapy research for both types of cancer. More-



(a) $\alpha = 0.1$ (b) $\alpha = 0.01$ (c) full dimensionality

Fig. 3. Venn diagrams with common edges to breast and prostate for regularised and full approaches, BNs learnt from tumour (T) and normal (NT) tissue samples.

over, this can be extended to multiple diseases, in the search for common gene signatures across multiple types of cancer.

References

1. G.P. Risbridger, I.D. Davis, S.N. Birrell, W.D. Tilley. "Breast and prostate cancer: more similar than different". *Nature Reviews Cancer*, vol.10, no.26, pp. 205-212, 2010.
2. P. McCullagh and J.A Nelder. "Generalized linear models". 2nd Edition, Chapman and Hall, London, 1989.
3. A.E. Hoerl and R.W. Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". *Technometrics*, vol.12, no.1, pp.55-67, 1970.
4. R. Tibshirani. "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society. Series B (Methodological)*, pp.267-288, 1996.
5. H. Zou and T. Hastie. "Regularization and variable selection via the elastic net". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no.2, pp.2601-320, 2005.
6. B. Aragam, Bryon, J. Gu, Q. Zhou. "Learning Large-Scale Bayesian Networks with the sparsebn Package". *arXiv preprint arXiv:1703.04025*, 2017.
7. The Cancer Genome Atlas - TCGA. https://cancergenome.nih.gov/
8. The Ensembl genome browser. Accessed May 2017, from https://www.ensembl.org/index.html
9. The Consensus CDS (CCDS) project. Release 20, Accessed May 2017, from https://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi
10. D. Szklarczyk, A. Franceschini et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, vol.43, pp.D447-52, 2015.