Journal of Applied Bioinformatics & Computational Biology

Research Article

A SCITECHNOL JOURNAL

Simple, Effective and Stochastic Method for Validation of Transcriptomic Data

Nicolas Nafati* and Samir Hamamah

Abstract

Background: Sequencing of the transcriptome has revolutionized quantitative and qualitative analyzes of prokaryotic and eukaryotic organisms. In the field of research on medically assisted reproduction, the selection of embryos with the best potential for implantation is the main challenge for biologists. Several studies suggest that the genes involved in oocyte cell cross talk could represent biomarkers for the selection of embryos with the greatest potential for implantation. Variability of different sources was observed during the transcriptomic experiment. Thus, one could have reasonable doubt about the validity of these transcriptomic data to provide a reliable and robust predictive model of pregnancy. Event is to be predicted at best by maximizing the likelihood criterion that will be discussed later. This vector Y which takes only two modalities: Positive or negative pregnancy is a random binary categorical and dependent process that follows a bi-nomial probability distribution. In the present study, the data are composed of 21 computative biomarker genes analyzed by quantitative polymerase chain reaction of 102 cumulus cell samples from patients undergoing *in vitro* fertilization.

Results: The Stochastic Likehood and the Youden's Index results will make it possible to discard the data in the case where they are biased and thus save time in terms of processing.

Conclusion: Indeed, the Stochastic Likehood and the Youden's Index showed that the transcriptomic data used are biased.

Keywords

Assisted Reproduction Technology (ART); Oocyte cumulus cells; Biomarkers; Stochastic Likelihood (SL)

Introduction

The data used here were obtained by analyzing a transcriptomic signature of 21 genes on 102 cumulus samples. Our goal is to optimize the time for data analysis and interpretation. Stochastic Likehood (SL) and Youden's Index (YI) are used to extract and visualize important information provided in a multivariate data table. This combination method provides decision information in terms of positive or negative pregnancy. In general, variables are often standardized. This is particularly recommended when variables are measured in different units (For example: Kilograms, kilometers, centimeters, otherwise, the result of the SL and YI obtained will be strongly affected. The goal is to make the variables comparable. Typically, the variables are normalized so that they ultimately have a standard deviation of one and an average of zero. Technically, the approach is to transform the data by substracting a reference value (the average of the variable) from each value and dividing it by the standard deviation. At the end of this transformation, the data obtained are called centric-reduced data. Data standardization is a widely used approach in the context of analyzing gene expression data prior to any pre-processsing analyzes [1-3].

When normalizing variables, the data can be transformed as follows:

(X-mean (X))/sd (X)

where mean (X) is the average of the values of X, and sd (X) is the standard deviation. The scale function can be used to normalize the data under R.

The formulation of the problem is that the prengancy variable Y to explain takes only two modalities: The positive or negative pregnancy (Pr^{+}/Pr^{-}), is a categorical and dependent binary random process that follows a binomial probability distribution with the parameters (N, p).

This random variable is formulated as follows:

$Y = X^{*} \beta^{T} + \mathcal{E}$

-X is the descriptive matrix $\in R^{N}*R^{p+1}$. This matrix is given by the following formula:

$$X = \begin{bmatrix} 1\, x_{11}\, \ldots.\, x_{1P} \\ \cdots\cdots\cdots\cdots.. \\ \cdots\cdots\cdots\cdots\cdots \\ 1\ x_{N1}\, \ldots.\, x_{NP} \end{bmatrix}$$

-β is the predictor vector $\in R^{p+1}$.

-Ɛ is the residual error following a normal distribution defined on the space Ω (0, σ2) of zero mean and of variance=σ2. The explanatory variable X represents the gene expression data. It is an explicative matrix of N rows (cumulative), and P genes. The goal is to predict the Y vector. Modeling the dependence of Y as a function of the descriptive matrix X is equivalent to estimating the optimal vector predictor β̂. This is done by maximizing the Log of the Likelihood. The objective of this optimization is to obtain the optimum predictor vector β̂. This by maximising the Log of Likelihood L_N (β) :

$$L_N(\beta) = \prod_{k=1}^{N} P_\beta(\{(Y_i = y_i | X_i = x_i\})$$

With $x=(x_1,x_2,\ldots,x_N)$ a statistical realization-sample of the transcriptomic matrix $X=(X_1,X_2,\ldots,X_N)$; the logistic model proposes a modelling of the law of Y | X=x by a Bernoulli distribution with parameter p_β (x) $=P_\beta$ ({(Y=1|X=x}) that represents the probability of observing the pregnancy event Y [4-7].

The Log of L_N(β) has to be maximised in order to compute the optimal β and then Y.

To obtain this Maximum, the Log(L_N (β)) is derived in order to obtain the optimal predictor vectorβ, and consequently the vector Odd-Ratio OR such that:

OR= exp(β^{T} X)

The interpretation of the coefficients β (Model) is generally done using Odds-Ratio (OR).

***Corresponding author:** Nicolas Nafati, Department of Biology, University of Montpellier, Montpellier, France, E-mail: nicolas.nafati@inserm.fr

Received: 06 April, 2020, Manuscript No. JABCB-20-8981; **Editor assigned:** 09 April, 2020, PreQC No. JABCB-20-8981 (PQ); **Reviewed:** 23 April, 2020, QC No. JABCB-20-8981; **Revised:** 10 August, 2022, QI No. JABCB-20-8981 (QI) Manuscript No. JABCB-20-8981 (R); **Published:** 07-September-2022, DOI: 10.37532/2329-9533.1000248.

The determination of the OR parameter is optimal when the likelihood leads us to an optimal predictor β and consequently to a model (targeted) also optimal.

We show that the Odd Ratio can be written as follows:

$$OR = \frac{p_\beta(x)}{1-p_\beta(x)} \Big/ \frac{p_\beta(\bar{x})}{1-p_\beta(\bar{x})}$$

We remark that the OR involving probabilities calculated according to the likelihood criterion. The general idea is to reason in terms of probability, using the ratings (Cote=Chance). The Odd-Ratio measures the relationship between the score of an event occurring in group of individuals A and the rating of the same event occurring in group of individuals B. The interpretation of the coefficients β (Model) is generally done using Odds-Ratio (OR). The determination of the OR parameter is optimal when the Likelihood leads us to an optimal predictor β and consequently to a model (targeted) also optimal [8].

The Youden's Index noted YI is also used such as:

YI= Se+Sp −1

This measures the accuracy of the diagnostic method. It depends on the Specificity (Sp) and Sensitivity (Se) of the test, but not on the prevalence of the event (Pr+/Pr-). The Youden's Index varies between -1 and 1. When equal to 0, it indicates that the test is ineffective. The diagnosis is maximal when the index YI is close to 1. It therefore allows the quantification of the informative value of a clinical sign, such as the pregnancy event [9].

Materials and Methods

Our method is based on the combined use of optimization of the Likehood Criteria and the use of the Youden's Indicator, which permet to draw these mentioned conclusions regarding the used data. Indeed the maximization of the Log of the likelihood criteria leads us to an optimal model (optimal predictor) and consequently to an optimal (OR).

The Log(L_N (β)) is derived in order to obtain the optimal predictor vector β, and consequently the optimal Odd-Ratio vector (OR) such that:

$$(OR) = \exp(\beta^T X)$$

The interpretation of the coefficients β (Model) is generally done using Odds-Ratio (OR). The Logit transformation makes the matrix system of optimization a linear system, and then making it possible to obtain the estimated pregnancy vector Y.

$$\mathrm{Logit}\big(p_\beta(x)\big) = \mathrm{Log}\left[\frac{p_\beta(x)}{1-p_\beta(x)^*}\right] = \beta_0 + \sum_{j=1}^{P=21} \beta_j x_j$$

This function (Logit) is bijective and differentiable, thus transforming the binary space (of Y to a real space)-∞+∞).

```
Algorithms

Initialization

DATA_1 (Learn_Table=2/3 of the raw data).

DATA_2 (Test_Table=1/3 of the raw data).

Extraction of binary vectors Y(Pr) and construction of Explicative Matrix X.

 Predictor Processing (under R)

    β_0  initialization.

    k ← 1

        Repeat

        $\beta^{k+1} \leftarrow \beta^k + A^k \nabla L_N(\beta^k)$ where $\nabla L_N(\beta^k)$ is the
        gradient at the point $\beta^k$ with $A^k$ is such that
        $A^k = (\nabla^2 L_N(\beta^k))^{-1}$ (called step matrix of the
        algorithm).

        k ← k + 1

    Until Convergence

    Predicted Probability Computation.

    Calculation of the (OR)= exp(β^T X)

    Calculation of the YI's
```

Results

Estimated pregnancy results

Here we give the estimated binary pregnancy vector (red curve) and the true pregnancy vector Y (black curve) (Figure 1) [10].

We note a difference between the measured and the estimated values of pregnancy.

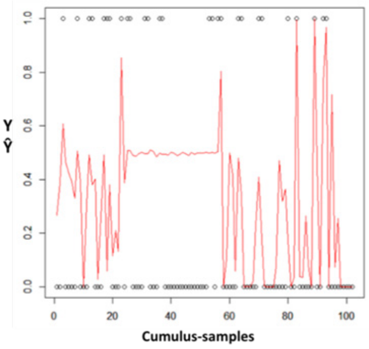

Figure 1: X axis=cumulus-samples. Y axis=estimated vecteur Ŷ (red curve) and the true pregnancy vector Y (black small circles).

We note a difference between the measured and the estimated values of pregnancy.

OR results

As indicated below we give the OR curves surrounded by the confidence intervals for all different genes (Figure 2).

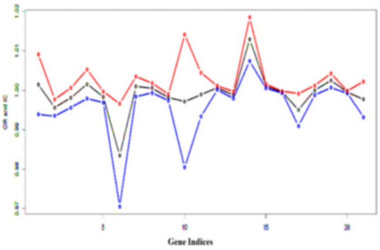

Figure 2: X axis=gene indices. Y axis=OR, black curve. The blue and the red curves correspond, respectively to the lower and the upper bound of the OR confidence intervals.

Note: The odds ratio is always greater than or equal to zero. If the odds ratio is close to 1, the event is independent regarding the group. This characteristic is given in the reference under the license "N ° CC BY-SA 3.0 Copyright Wikipedia author (s)".

The OR should reach its optimum when the model to be studied reaches its own optimum. We observe that the OR is equivalent on average to 1 practically for all genes, which corresponds to the absence of the event Pr [11].

YI's results

The original input data is characterized by their results in which the step sampling is: First half data as the leaning data, the other half is as the validation data. For as threshold of 0.5, we find these results [12].

-Sensitivity-Se=0.4444-Specificity-Sp=0.8307-YI=0.2751

Then, the step sampling is the first 2/3 of data, it is used as leaning data, the other 1/3 of data is for the validation and at the same threshold:

-Sensitivity-Se=0.7286-Specificity-Sp=0.2581-IY=-0.0133

Discussion

True and estimated values of the vector Y are widely different in the sense of likelihood. We can say that the analysis of OR (OR= exp(β^T X)) indicates a low informative contribution in terms of diagnostic effectiveness. The YI is far from 1, which means that the informative value of the clinical sign is low. The Likelihood translating by the Odd Ratio, the Youden's index are powerful tools for the analysis and interpretation of biological tests in terms of diagnosis [13,14].

It can therefore be deduced from the obtained results that these transcriptomic data do not allow obtaining a model that discriminates in the absolute sense, consequently, these data are not exploitable [15].

Conclusion

It therefore appears that these data are highly correlated indicating a single group of individuals whose variance is practically collinear with axis 1 and more or less representative since cos2 is less than or equal to 1. Finally, we say that the analysis of OR which contains the optimal predictor that is calculated according to the criterion of the Linklihood means that the informative value of the clinical sign is low.

There is a big difference between the observed and the predicted pregnancy event. This reinforces the doubt about the model validity. Finally we say that the combination of the Likelihood criteria (translating by the computation of OR), the Youden's Index are powerful tools for biological test's analysis and interpretation.

Ethics Approval and Consent to Participate

The Review Board of the University Hospital approved this project. All patients provided their written informed consent for the use of cumulus samples for research.

I.R.M.B Monpellier University Hospital. Accredidation Number: 1987//

Consent for Publication

I declare that I fully authorize the publication, reproduction, dissemination and any other use of my scientific article.

Availability of Data and material

NA.

Computing interests

No competing interests.

Funding

INSERM.

Author contribution

NN and SH conceived the study. NN performed the mathematical analysis and wrote the manuscript.

Acknowledgements

Thanks are due to Ounissa Aït Ahmed for reading the manuscript.

References

1. Akaike H (1973) Information theory and an extension of the maximum likelihood principle, proceedings of the 2nd international symposium on information.

2. Bagley SC, White H, Golomb BA (2001) Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol 4:979-985.

3. Bardos M (2001) Analyse discriminante: Application au risque et scoring financier.

4. Birch DE, Kolmodin L, Wong J, Zangenberg GA, Zoccoli MA, et al. (1996) Simplified hot start PCR. 381:445-446.

5. Habibzadeh F, Habibzadeh P (2019) The likelihood ratio and its graphical representation. Biochem Med 29:193-199.

6. Heid CA, Stevens J, Livak KJ, Williams PM (1996) Real time quantitative PCR. Genome Res 986-994.

7. Lebart L, Morineau A, Piron M (1995) Statistique exploratoire multidimensionnelle.

8. Konishi S, Kitagawa G (2000) Information criteria and statistical modeling.

9. McGee S (2002) Simplifying likelihood ratios. J Gen Intern Med 647-650.

10. Nendaz MR, Perrier A (2004) Sensibilité, spécificité, valeur prédictive positive et valeur prédictive négative d'un test diagnostique. Revue des maladies respiratoires 21:390-393.

11. Schisterman EF, Perkins NJ, Liu A, Bondell H (2005) Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. Epidemiol 73-81.

12. Thornbury JR, Fryback DG, Edwards W (1975) Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information. Radiol 114:561-565.

13. Mellins HZ, McNeil BJ, Abrams HL, van Houten FX, Murphy MA, et al. (1979) The selection of patients for excretory urography. Radiol 130:293-296.