Estimation of Gynecological Cancer Networks via Target Proteins

Başak Bahçivancı
Department of Statistics
Middle East Technical University
Ankara, Turkey

Eda Purutçuoğlu
Department of Social Services
Ankara University
Ankara, Turkey

Vilda Purutçuoğlu
Department of Statistics
Middle East Technical University
Ankara, Turkey

Yüksel Ürün
Department of Medical Oncology
Ankara University School of Medicine
Ankara, Turkey

Abstract*—The construction of biological networks has certain challenges due to its high dimension, sparse structure and very limited number of observations. Thus, specific modeling approaches have been suggested to deal with these problems such as Gaussian graphical model, loop-based multivariate adaptive regression splines (MARS) with/without interaction effects and Gaussian copula graphical model. From previous analyses via these methods, it has been shown that they can successfully estimate the systems with comparative accuracies. Hereby, in this study, as the novelty we use all these complex mathematical models in inference of gynecological cancer networks whose target genes are gathered from biological literature. The observations for these target genes are collected from the ArrayExpress database with other associated risk factors such as stage of the cancers which are denoted by categorical variables. Then, under different dimensions of systems, sample sizes and measurement types, we compare the performance of all models with the criteria of accuracy and F-measure. From the results, we observe that the suggested models can successfully estimate the real cancer systems under different conditions and are promising approaches to describe the complexity in biological networks.

***Keywords*—mathematical models; biological networks; gynecological cancers**

I. Introduction

Cancer is a branch of systems’ diseases which are caused by malfunctions in different regulation pathways affecting cellular proliferation, growth and apoptosis. The gynecologic cancer is one group in these illnesses which have influence upon women’s reproductive organs which cover vaginal, vulvar, ovarian, cervical and endometrial parts. After breast cancer, gynecological cancers are still the second most common cancer types which affect approximately 1:4 of all cancers in women in developing countries (excluding non-melanoma skin cancer) [1]. Despite the fact that gynecological cancers' incidence decreases in some countries, whereas in others like less developed countries, the incidence rate rises during the same interval [2]. Therefore, accurately screening of potentially crucial genes and their interactions, which could pave the way for significant breakthroughs in finding cures for various cancers like gynecological cancers, become more of an issue. However, there are challenges to understanding the underlying system model due to the functional and structural complexity inherent in biological systems [3, 4]. For example, the high sparsity of the system and the high correlation between the genes are the two of the several issues that should be considered about inference of the gene regulatory systems. Therefore, in order to derive a plausible interaction network and to make a reliable inference about networks to explain an actual system's disease, the choice of mathematical models plays a significantly crucial role [5, 6].

Hereby, in this study, we initially make a list of quasi target proteins of these cancers by combining various oncogene researches [7, 8, 9] and we generate a target pathway having 11 genes whose biological interactions are also validated by the literature. Then, we check the ArrayExpress database and find all the Affymetric data which include the underlying genes. Finally, we perform various complex mathematical models to estimate the pathway and control the accuracy of the models. By this way, we aim to find the best modelling approach which can be used for the construction of the actual pathway disease.

Accordingly, we present the recently developed networks models which are particularly applied for high dimensional and sparse biological networks in Section 2. Then, we introduce our datasets in Section 3. In Section 4, we report the application of these models to our data and tabulate the accuracy measures of our models for each dataset. Lastly, in Section 5, we summarize our findings and discuss the future works.

II. Mathematical Models

There are a number of mathematical models designed for the description of biological networks under various assumptions. In this study, we perform GGM, CGGM with RJMCMC and BDMCMC inference methods as well as the LMARS model with/without gene interaction effect. Below, we present each alternative shortly.

A. *Gaussian Graphical Model (GGM)*

This model is one of the fundamental modeling approaches in the steady-state activation of the system via the undirected graph. Under the assumption of multivariate normality, GGM uses the following model to regress each gene in the system on the remaining genes.

$$Y_p = \beta Y_{-p} + \varepsilon , \tag{1}$$

where Y_p and Y_{-p} denote the state of the *p*th state and all the remaining states except *p*, respectively. Moreover, β and ε show the regression coefficient and the random error, in order. In Equation 1, the state Y is assumed to have a multivariate normal distribution with mean μ and variance Σ and there is a direct relation between the inverse of Σ, denoted by θ and called the precision matrix, and β which is the model parameter in (1). In biological networks, θ is sparse, meaning that there are many zeros in the entry of θ. Accordingly, the relation between θ and β is described by the following expression.

$$\beta = -\frac{\theta_{-p,p}}{\theta_{p,p}} \tag{2}$$

in which $\theta_{-p,p}$ shows the precision when the *p*th state is excluded and $\theta_{p,p}$ indicates the precision with totally *p* genes. Due to the feature of the multivariate normality, the zero value in θ implies the conditional independence between the associated pair of genes and the non-zero entry represents the existence of the interrelation between those genes. Therefore, in GGM, we prefer to infer θ, rather than β, in order to use this special feature of the conditional independency. Hereby, to estimate θ in GGM, various approaches are suggested. Graphical lasso [10], neighborhood selection [11], adaptive lasso and fused lasso are some examples. But among them, graphical lasso, which is based on the L_1 – penalized likelihood equation, is the most common one, thereby, is chosen in this study too.

B. *Copula Gaussian Graphical Model (CGGM)*

When the dimension of the systems increases, the GGM cannot estimate the systems very accurately since the inference is based on the likelihood equation and the sparsity in the systems becomes a challenge in inference. Therefore, GGM is redesigned by changing the joint function of the multivariate normality of states by partitioning it via the Gaussian copula. This new representation of the high dimensional joint distribution enables us to estimate θ via the Bayesian algorithms which are free from the limitation of the high dimension and the problem of the sparsity [12, 13].

There are two major alternative inference methods for CGGM. These are the reversible jump Markov chain Monte Carlo (RJMCMC) approach [12, 13] and the birth-and-death MCMC (BDMCMC) method [14]. The former is the modified version of the Metropolis-Hasting algorithm which provides jumps between spaces of different dimensions under three stages whose mathematical details can be found in Dobra and Lenkoski (2011) [12]. On the other hand, the second method is based on the two poisson rates for the acceptance and the rejection of links between genes in a MCMC scheme whose mathematical details can be seen in Mohammadi and Wit (2015) [14]. Besides these approaches, theoretically, there are other alternative Bayesian inference methods that can be adapted in CGGM. The split and merge approach [15], Carlin and Chibs method [16] and the Gibbs sampling [17, 18] are some of these examples. In this study, we perform CGGM under RJMCMC and BDMCMC due to the fact that their adaptations for sparse systems have been discussed comprehensibly and their R programmes which simplify their applications have been already developed.

C. *Loop-based Multivariate Adaptive Regression Splines (LMARS)*

The MARS model is one of the well-known models in the family of generalized additive models [19, 20]. MARS is specifically designed for highly dependent data under nonlinear structures. The full description of this model is presented as below.

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(x) + \varepsilon , \tag{3}$$

where β_0 denotes the intercept, β_m stands for the regression coefficient and ε implies the random error, as used previously. Here, *M* shows the total number of parameters and B_m is the basis function which has the following form.

$$B_m(x) = \prod_{k=1}^{K_m} \left[\max\left(s_{k,m} (x_{v(k,m)}) - t_{k,m}, 0 \right) \right] \tag{4}$$

in which K_m and $s_{k,m}$ present the number of truncated linear functions multiplied in the *m*th basis function and the input variable corresponding to the *k*th truncated linear function in the *m*th basis function, respectively. In this expression, the basis function is chosen via the condition below.

$$(x-t)_+ = \begin{cases} x-t & if \quad x>t \\ 0 & otherwise \end{cases},$$

$$(t-x)_+ = \begin{cases} t-x & if \quad x<t \\ 0 & otherwise \end{cases}.$$

In the expression above, *t* is the knot which presents the breaking point of the spline function. The knot is used to get the reflected pairs for each x_j with knots at each observed value x_{ij} .

In the application of this model in biological networks, we need to eliminate very high orders of interactions since there is no biological correspondence of such relationships in the literature of systems biology. Thereby, we convert (3) into (1) so that the biological systems can be nonparametrically represented by MARS similar to the GGM approach. This revised model is called the loop-based MARS (LMARS) [21] which is based on solely the main effects. Later, we also extend this LMARS without interaction model by adding merely the second order interaction effects [22]. The second order interaction implies the feed forward loop [23] in biological systems and we call this model as "LMARS with interaction" in this study.

III. Data Description

In the construction of a quasi-gynecological network, we firstly check the related literature about these cancers which mainly cover cervix, ovarian and endometrial cancers. Then, we detect that CTNNB1, TFAM, CEBPM, MAP2K1, MAPK1, TP53, PDIA3, IMP3, ERBB2, CHD4 and MBD3 are the core genes which have direct influence in these cancers [7, 8, 9] and have dense connections between each other in such a way that they generate a complete graph. This means that all genes have connections with the remaining genes, resulting in an adjacency matrix made up of fully "one" entries. Then, we comprehensively control the ArrayExpress dataset to find the gene expression data which consist of these 11 genes. During the data collection, we choose the ArrayExpress database [24] since it is one of the most common databases which is composed of high throughout functional genomics data. This dataset also supports community standards such as MIAME and MAGE-ML [25, 26]. Hereby, we present the biological description of each dataset which has our 11 genes as below.

i) E-GEOD-9891 - Transcription Profiling of 285 Human Ovarian Tumour Data: The data are a cohort of 285 patients with epithelial ovarian, primary peritoneal, or fallopian tube cancer, diagnosed between 1992 and 2006. They are identified through Australian Ovarian Cancer Study8 (sample size n = 206), Royal Brisbane Hospital (n = 22), Westmead Hospital (n = 54) and Netherlands Cancer Institute (NKI-AVL; n = 3) [27]. In this dataset, the arrays are designed by randomly selected samples from the Australian Ovarian Cancer Study whose expression profiles on the Affymetrix U133_plus2 platform aim to identify novel subtypes of the ovarian tumour by gene expression profiling with linkage to clinical and pathologic features [27].

ii) E-GEOD-63678 – Expression Data from Vulvar, Cervical, Endometrial, Carcinoma Tissue: In this dataset, 35 samples are used to identify potential biomarkers and signatures in each type of cancer. Specifically, 18 cancer samples with 5 cervical, 7 endometrial and 6 vulvar cancers, and also 17 normal samples with 5 cervical, 5 endometrial and 7 vulvar cancers are hybridized on the Affymetrix platform in order to identify the common features among cancer types, embryonic stem cells and the newly discovered cell population of squamocolumnar junction of the cervix, considered to host the early cancer events [28]. Moreover, total RNA is extracted from physiological and cancer patients from cervix, endometrium and vulvar tissue and is hybridized on the Affymetrix HG133_A_2.0 microarray chips corresponding to more than 12.000 uniquely represented genes [28].

iii) E-GEOD-81248 - Expression Data from HEY Cells: The data are collected from two samples of distal naïve cells that are exposed to either unstimulated (control) and stimulated (LPS or poly(I:C)) exosomes from local cells in order to show whether the exosomes from TLR stimulated cells can largely recapitulate the TLR activation in distal cells in vitro [29]. In this dataset, the mRNA expression is captured on the Affymetrix U133 Plus 2 chips and includes all the 11 core genes which we list previously. In the dataset, each gene has 12 observations.

iv) E-GEOD-14764 - A Prognostic Gene Expression Data in Ovarian Cancer: The data are a cohort of 80 ovarian carcinomas (TOC cohort) for the development of a predictive model, which is then evaluated in an entirely independent cohort of 118 carcinomas (Duke cohort) [30]. In this dataset, RNA from 80 frozen ovarian cancer samples is hybridized on Affymetrix Human Genome U133A Array and the collected data contain our 11 core genes. In the data collection, it is aimed to investigate the hypothesis that molecular markers are able to predict outcome of the ovarian cancer independently from classical clinical predictors, and that these molecular markers can be validated using independent datasets [30].

IV. Application

D. *Accuracy Measures*

In our analyses, in order to compare the performance of methods, we use the accuracy value and F-measure for all datasets. These measurements are calculated with the following four main values: true positive, true negative, false positive and false negative value. True positive (*TP*) indicates the number of correctly classified objects that have a positive label. The true negative (*TN*) implies the number of correctly classified objects that have a negative label. On the other hand, the false positive (*FP*) presents the number of misclassified objects that have a negative label and the false negative (*FN*) shows the number of misclassified objects that have a positive label. This information also constructs a confusion matrix as shown in Table 1, which represents the actual and the predicted classification.

TABLE I. General confusion matrix.

		Actual	**class**
		Positive	**Negative**
Predicted class	**Positive**	TP	FP
	Negative	FN	TN

Hereby, the accuracy is expressed by the ratio of correctly classified objects in both labels to the total of all classified objects which is shown by the following formula:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}. \qquad (5)$$

On the other hand, F-measure is calculated by the equation below.

$$F-measure = 2\frac{Precision \times Recall}{Precision + Recall}, \qquad (6)$$

where

$$Presicion = \frac{TP}{TP+FP} \text{ and } Recall = \frac{TP}{TP+FN}. \quad (7)$$

In the following part, we present the performance of each model for every dataset separately and evaluate the findings.

E. *Results of Datasets*

In the following tables, we tabulate the accuracy of each suggested model for the selected datasets. In the calculation, we compare the estimated networks with the true system. This system is composed of 11 genes and all the interactions between these genes are validated via the STRING database. Hereby, the true network with these 11 genes shows a complete graph which implies fully one in its adjacency matrix.

As seen in Table 2, the best accuracy is obtained in the CGGM approach under the parametric models and LMARS with interaction approach under the non-parametric models when we have small dimensional networks.

Then, in order to see specifically the effect of the increase in dimensions on the accuracy of models, we extend the study by using 1000 genes for each dataset including the core 11 genes. In other words, in the analyses, we still check the accuracy of links in the same 11 genes, but this time, we evaluate the models under a high dimensional setting which is composed of 1000 genes in the system. Hereby, we tabulate our findings in Table III for E-GEOD-9891, E-GEOD-63678 and E-GEAD-14764 datasets. In these calculations, E-GEOD-81248 is not applied since there is less number of genes in that dataset. Furthermore, we discard the results of CGGM too since both RJMCMC and BDMCMC estimation methods become computationally very demanding when the dimension of the systems is extended to 1000 genes. Therefore, in Table III, we present the outcomes of GGM, LMARS with/without interaction effect models.

TABLE II. comparison of F-measure and accuracy for all datasets.

DATA: E-GEOD-9891		
METHODS	**F-MEASURE**	**ACCURACY**
GGM	0.167	0.091
CGGM VIA RJMCMC	**0.846**	**0.733**
CGGM VIA BDMCMC	0.448	0.289
LMARS WITHOUT INTERACTION	0.752	0.603
LMARS WITH INTERACTION	**0.858**	**0.752**
DATA: E-GEOD-63678		
METHODS	**F-MEASURE**	**ACCURACY**
GGM	0.167	0.091
CGGM VIA RJMCMC	**0.981**	**0.791**
CGGM VIA BDMCMC	**0.964**	**0.655**
LMARS WITHOUT INTERACTION	0.726	0.570
LMARS WITH INTERACTION	0.778	0.636
DATA: E-GEOD-81248		
METHODS	**F-MEASURE**	**ACCURACY**
GGM	0.091	0.167

CGGM VIA RJMCMC	**0.472**	**0.309**
CGGM VIA BDMCMC	**0.736**	**0.582**
LMARS WITHOUT INTERACTION	0.429	0.273
LMARS WITH INTERACTION	0.408	0.256
DATA: E-GEOD14764		
METHODS	**F-MEASURE**	**ACCURACY**
GGM	0.167	0.091
CGGM VIA RJMCMC	**0.911**	**0.836**
CGGM VIA BDMCMC	0.226	0.127
LMARS WITHOUT INTERACTION	0.193	0.107
LMARS WITH INTERACTION	**0.752**	**0.603**

TABLE III. COMPARISON OF F-MEASURE AND ACCURACY FOR THREE DATASETS UNDER 1000 GENES' SYSTEMS.

DATA: E-GEOD-9891		
METHODS	**F-MEASURE**	**ACCURACY**
GGM	0.271	0.157
LMARS WITHOUT INTERACTION	0.246	0.140
LMARS WITH INTERACTION	**0.271**	**0.157**
DATA: E-GEOD-63678		
METHODS	**F-MEASURE**	**ACCURACY**
GGM	0.167	0.091
LMARS WITHOUT INTERACTION	**0.193**	**0.107**
LMARS WITH INTERACTION	0.167	0.091
DATA: E-GEOD-14764		
METHODS	**F-MEASURE**	**ACCURACY**
GGM	0.167	0.091
LMARS WITHOUT INTERACTION	**0.193**	**0.107**
LMARS WITH INTERACTION	0.049	0.025

From the results in Table III, it is seen that the accuracy of all models decreases when the dimension of the systems increases. Moreover, the best performance is observed under the LMARS without interaction models.

In the final stage of the analyses, we use a single dataset (E-GEOD-9811) since it is the only set which has discrete measurements besides its microarray observations. In this dataset, the following categorical variables are also publically available and they are included in modeling of the system as the risk factors of the disease [5].

- Primary site: composed of three stages, namely, ovary, peritoneum and fallopian.
- Type: consists of two groups, namely, malignan and LMP.
- Subtypes: consisting of 3 groups, namely, ser/papser, endo and adeno.
- Consolidated grade: recorded under three levels, namely, 1, 2 and 3.

The tabulated values are shown in Table IV for GGM and LMARS with/without interaction effect models. In these analyses, since GGM can work with solely continuous measurements and cannot deal with categorical datasets, we add jitters to those listed categorical risk factors in order to convert all measurement to continuous scales. On the other hand, we do not make any adjustments for both LMARS models as they can be used for both discrete/categorical and continuous datasets. Finally, in our comparison, we still apply 1000 genes in modeling. From the results in Table IV, it is found that the best accuracy is computed for the LMARS with interaction model and the performance of this model is improved significantly when the model is extended via the categorical risk factors.

TABLE IV. COMPARISON OF F-MEASURE AND ACCURACY FOR E-GEOD-9891 DATA UNDER 1000 GENES.

DATA: E-GEOD-9891		
METHODS	**F-MEASURE**	**ACCURACY**
GGM	0.208	0.116
LMARS WITHOUT INTERACTION	0.542	0.372
LMARS WITH INTERACTION	**0.683**	**0.521**

V. CONCLUSION

In this study, we have aimed to generate a gynecological cancers pathway by choosing the core genes in this disease. For this purpose, we have combined the biological literature about this illness and generate a list of genes composed of 11 core proteins. Then, we have searched all microarray datasets in the ArrayExpress database which has our 11 core species. We have found 4 datasets under different sample sizes per gene. Then, we have modelled them separately by the selected approaches, which are Gaussian graphical model (GGM), copula GGM with two different inference approaches and the loop-based multivariate adaptive regression splines (LMARS) model constructed by both gene effect and second order gene-interaction effect conditions. Here, we have chosen these modelings since they are particularly designed for the construction of complex biological networks. In our analyses, we have evaluated the performance of these 4 mathematical descriptions in a real disease network. In our calculation, we have assessed the models under moderate and high dimensions as well as under continuous and discrete measurements. The results have indicated that for the small dimensional systems, the CGGM model has a better performance. Whereas, since it is computationally very demanding, it has a limited application. On the other hand, for high dimensions, LMARS, particularly, with the gene–interaction effect version is more successful than alternates. When we have applied both discrete and continuous measurements, it has been seen that still LMARS can accurately construct the systems.

As a result, we have concluded that among parametric models CGGM can be the most promising approach and among nonparametric models, LMARS with interactions (i.e., gene-interaction effects) is more preferable in real-life datasets. As the extension of this study, we consider to collect more datasets from different databases and evaluate the performance of these models. We consider that this study can be a first step to generate a target drug in personalized medicine for the gynecological cancers' studies and can be adapted to other complex systems illnesses such as heart diseases and other cancers types.

Acknowledgment

The authors would like to thank to the Middle East Technical University research grant (no: BAP-08-11-2017-035) for its support. Moreover, they gratefully thank to the anonymous referees and the editor-in-chief for their recommendations.

References

[1] C. A. Iyoke and G. O. Ugwu, "Burden of gynaecological cancers in developing countries", World Journal Obstetrics Gynecology, 2 (1), pp.1-7. 2013.

[2] S. Pecorelli, G. Favalli, L. Zigliani, and F. Odic, "Cancer in women", International Journal of Gynecology and Obstetrics, 82 (3), pp. 369–379. 2013.

[3] G. W. Carter, C. G. Rush, F. Uygun, N. A. Sakhanenko, D. J. Galas, and T. Galitski, "A systems-biology approach to modular genetic complexity", Chaos, 20 (2), pp. 026102.1-8. 2010.

[4] A. Noor, E. Serpedin, M. Nounou, H. Nounou, N. Mohamed, and L. Chouchane, "An overview of the statistical methods used for inferring gene regulatory networks and protein-protein interaction networks", Advances in Bioinformatics, 953814. 2013.

[5] J. M. Bower, and H. Bolouri, Computational Modeling of Genetic and Biochemical Networks. MIT Press, Cambridge. 2001.

[6] D. J. Wilkinson. Stochastic Modelling for Systems Biology. Taylor and Francis, Boca Raton, FL. 2006.

[7] The Cancer Genome Atlas Research Network, "Integrated genomic analyses of ovarian carcinoma", Nature, 474, pp. 609–615. 2011.

[8] Cancer Genome Atlas Research Network, "Integrated genomic characterization of endometrial carcinoma", Nature, 497, pp. 67–73. 2013.

[9] Z. Hu, D. Zhu, W. Wang, W. Li, W. Jia, X. Zeng, and et al., "Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration", Nature Genetics, 47, pp. 158-163. 2015.

[10] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso", Biostatistics, 9, pp. 432-441. 2008.

[11] N. Meinshausen, and P. Bühlmann, "High dimensional graphs and variable selection with the lasso", The Annals of Statistics, 34, pp. 1436-1462. 2006.

[12] A. Dobra, and A. Lenkoski, "Copula Gaussian graphical models and their application to modeling functional disability data", Annals of Applied Statistics, 5, pp. 969-993. 2011.

[13] H. Farnoudkia, and V. Purutçuoğlu, "Copula Gaussian graphical modelling of biological networks and Bayesian inference of model parameters". Scientia Iranica (in press).

[14] A. Mohammadi, and E. C. Wit, "BDgraph: Bayesian structure learning of graphs in R", Bayesian Analysis, 10, pp. 109-138. 2015.

[15] S. Richardson, and J. Green, "On Bayesian analysis of mixtures with an unknown number of components", Journal of Royal Statistical Society, Series B, 59, pp. 731-792. 1997.

[16] B. P. Carlin and S. Chibs, "Bayesian model choice via Markov chain Monte Carlo methods", Journal of Royal Statistical Society, Series B, 57 (3), pp. 473-484. 1995.

[17] S. Walker, "A Gibbs sampling alternative to reversible jump MCMC", Electronic Journal of Statistics, arXiv:0902.4117, pp. 1-3. 2009.

[18] V. Purutçuoğlu and H. Farnoudkia, "Gibbs sampling in inference of copula Gaussian graphical model adapted to biological networks", Acta Physica Polonica, Series A, 132, pp. 1112-1117. 2017.

[19] T. Hastie, R. Tibshirani, and J. H. Friedman, The Element of Statistical Learning, Springer-Verlag, New York. 2001.

[20] J. H. Friedman, "Multivariate adaptive regression splines", Annals of Statistics, 19 (1), pp. 1-67. 1991.

[21] E. Ayyıldız, M. Ağraz, and V. Purutçuoğlu, "MARS as the alternative approach of Gaussian graphical model for biochemical networks", Journal of Applied Statistics, 44 (16), pp. 2858-2876. 2017.

[22] M. Ağraz and V. Purutçuoğlu, “Extended lasso-type MARS (LMARS) model in the description of biological network”, Journal of Statistical Computation and simulation, 89 (1), pp. 1-14. 2019.

[23] U. Alon, An Introduction to Systems Biology: Design Principles of Biological Circuits, Chapman and Hall/CRC, Boca Raton, FL. 2007.

[24] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, and B. Brazma, “ArrayExpress—a public database of microarray experiments and gene expression profiles”, Nucleic Acids Research, 35, pp. D747-D750. 2007.

[25] M. Schena, Microarray Analysis, John Wiley and Sons, Hokoben. 2003.

[26] E. Wit and J. McClure, Statistics for Microarray Design, Analysis, and Inference, 1st edn, John Wiley and Sons Ltd. 2004.

[27] R. Tothill, A. Tinker, J. George, R. Brown, S. Fox, D. Johnson, and et al., “Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome”, Clinical Cancer Research, 14 (16), pp. 5198-208. 2008.

[28] K. I. Pappa, A. Polyzos, J. Jacob-Hirsch, N. Amariglio, G. D. Vlachos, D. Loutradis, and N. P. Anagnou, “Profiling of discrete gynecological cancers reveals novel transcriptional modules and common features shared by other cancer types and embryonic stem cells”, PLoS ONE, 10 (11), pp. e0142229.1-20. 2015.

[29] S. Srinivasan, M. Su, S. Ravishankar, J. Moore, P. Head, J. B. Dixon, and F. Vannberg, “TLR-exosomes exhibit distinct kinetics and effector function”, Scientific Reports, 7, pp. 41623. 1-14. 2017.

[30] C. Denkert, J. Budczies, S. Darb-Esfahani, B. Györffy, J. Sehouli, D. Könsgen, R. Zeillinger, W. Weichert, A. Noske, A. C. Buckendahl, B. M. Müller, M. Dietel, and H. Lage, “A prognostic gene expression index in ovarian cancer—validation across different independent data sets”, The Journal of Pathology, 218, pp. 273–280. 2009.