RESEARCH ARTICLE

Bayesian Survival Analysis of High-Dimensional Microarray Data for Mantle Cell Lymphoma Patients

Azam Moslemi[1], Hossein Mahjub[2]*, Massoud Saidijam[3], Jalal Poorolajal[2], Ali Reza Soltanian[2]

Abstract

Background: Survival time of lymphoma patients can be estimated with the help of microarray technology. In this study, with the use of iterative Bayesian Model Averaging (BMA) method, survival time of Mantle Cell Lymphoma patients (MCL) was estimated and in reference to the findings, patients were divided into two high-risk and low-risk groups. Materials and Methods: In this study, gene expression data of MCL patients were used in order to select a subset of genes for survival analysis with microarray data, using the iterative BMA method. To evaluate the performance of the method, patients were divided into high-risk and low-risk based on their scores. Performance prediction was investigated using the log-rank test. The bioconductor package "iterativeBMAsurv" was applied with R statistical software for classification and survival analysis. Results: In this study, 25 genes associated with survival for MCL patients were identified across 132 selected models. The maximum likelihood estimate coefficients of the selected genes and the posterior probabilities of the selected models were obtained from training data. Using this method, patients could be separated into high-risk and low-risk groups with high significance (p<0.001). Conclusions: The iterative BMA algorithm has high precision and ability for survival analysis. This method is capable of identifying a few predictive variables associated with survival, among many variables in a set of microarray data. Therefore, it can be used as a low-cost diagnostic tool in clinical research.

Keywords: Survival analysis - Microarray - Bayesian model averaging - mantle cell lymphoma patients

Asian Pac J Cancer Prev, **17 (1)**, 95-100

Introduction

Among different types of cancers, the lymphoma are group of malignant diseases with the origin of lymphocytes (Fisher, 2005; Adamson et al., 2007). The lymphoma is divided into two types, the Hodgkin's lymphoma and the non-Hodgkin's lymphoma. Specific type of the non Hodgkin lymphoma is the Mantle Cell Lymphoma (MCL). The Mantle Cell Lymphoma contains 6% of all the non-Hodgkin lymphoma and a larger fraction of deaths from lymphoma (Campo et al., 1999; Swerdlow and Williams, 2002). In this type of cancer, length of survival after diagnosis is varied and many attempts have been done for prediction of survival these patients (Rosenwald et al., 2003). Microarray technology and access to thousands of gene expression have been helped to predict survival of patients (Rust et al., 2005; Li et al., 2007). But the main challenge of this method is that the number of variables (p) is usually too much more than number of samples (n) (p>>n). In this case, identifying the strongest gene variables determinant of patients survival is very difficult (Li et al., 2007).

On the other hand, the model-building process for survival analysis based on the gene expression, involves the comparison many competing models. In this method, a single model is choosing each time and a statistical inference is accomplished based on that. But, uncertainty of single model selection causes overall uncertainty of the desired quantity estimation and considerable reduction of fitting model (Kass and Raftery, 1993). One of the solutions that is presented to this problem, is using the iterative Bayesian Model Averaging.

The iterative Bayesian Model Averaging method, estimates and reports the effect of multiple models, by computation the weighted average of their posterior distributions instead of selecting a single model and to perform the statistical inference based on that (Annest et al., 2009). In present study, iterative Bayesian Model

*[1]Department of Biostatistics & Epidemiology, [2]Research Center for Health Sciences and Department of Biostatistics & Epidemiology, School of Public Health, Hamadan University of Medical Sciences, [3]Research Center for Molecular Medicine, Department of Molecular Medicine and Genetics, School of Medicine, Hamadan University of Medical Sciences, Hamadan, Iran *For correspondence: mahjub@umsha.ac.ir*

Averaging method, using the survival of the Mantle Cell Lymphoma patients is estimated and based on that, the patients were divided into two groups: high-risk and low-risk.

Materials and Methods

In this study, dataset related to 92 the Mantle Cell Lymphoma patients (MCL) has been used. Patients were included 72 (78%) males and 19 (22%) females. Their median age at diagnosis time was 61.5 years (range age of 38 to 92.5 years). Follow up continued to death time for all patients. For each patient, the 8810 genes, the survival time and censoring status have been measured. Survival time was different from 0.228 to 168.636. Median survival was 2 years and 7 months. Among 92 patients under investigation, 28 patients (30.4%) have been censored. This dataset published by Rosenwald et al and is available at http://llmpp.nih.gov/MCL (Rosenwald et al., 2003).

In order to, the fitting survival analysis model and its evaluation, overall dataset was considered as "train sample" and of their set 31 sampels randomly was selected as "test sample". The iterative Bayesian Model Averaging method performed on train sample. Obtained estimations were used for the assessment method in the test sample.

The strength of BMA method is to be able to estimate model uncertainty. Calculation model uncertainty is the part of Baysian model analysis that largely has been ignored by traditional stepwise selection methods. BMA solves this problem with selection a subset of all possible models and the statistical inference based on the weighted average of posterior distribution of models (Annest et al., 2009). In this method, Uniform distribution is considering as the prior distribution. The core of the BMA algorithm is described in Equation (1) (Raftery, 1995). Let Ψ and TD indicate desired quantity and train data, respectively. Let $S=\{M_1, M_2, \ldots, M_n\}$ represent the subset of selected models that enter in the analysis.

We have:

$$pr(\psi|TD)=\Sigma_{i\in S}\, pr(\psi|TD,M_i)\, pr(M_i|TD) \qquad (1)$$

Three objects must be considered before Equation (1) could be used:

i). Obtaining the subset S of models that enter in the analysis, *ii)*. Estimation the value of $pr(\psi|TD, M_i)$, *iii)*. Estimation the value of $pr(M_i|TD)$.

One of BMA's problems is number of models which potentially can be discovered by the algorithm, especially when working with microarray data. Two algorithms "leaps and bounds" and "Occam's window" are being used for solving this problem and discarding non-helpful models (Annest et al., 2009).

The leaps and bounds algorithm regards the best expected researcher number of the models (nbest) and based on that the top nbest models estimates with every number of variables (maximum 25 variables) (Furnival and Wilson, 1974). With using the Occam's window algorithm every model which its posterior distribution of the best model is less than the cut point determined by the researcher could be removed (Madigan and Raftery, 1994). It proposes to remove the models with a probability of less than 5% as likely as the strongest model (Annest et al., 2009). Thus the remaining models constitute the set S in Equation (1).

Estimation second object and accurate calculation of predictive distribution pr $(\psi|TD, M_i)$ requires integration of the vector of regression parameters θ_i, however integration is impossible for the most censored survival models, therefore maximum likelihood estimation (MLE) is used as an approximation:

$$pr(\psi|TD, M_i) \approx pr\,(\psi|\theta_i, TD, M_i) \qquad (2)$$

For estimation third object and calculation of the posterior probability of model M_i given the training data, involves an Integral. But, for this reason that is impossible to assess exact value third object, the Bayesian information criterion (BIC) can be used as approximation estimation:

$$\log[pr(TD|M_i)]=\log[pr(TD|\theta_i, M_i)]-(\tfrac{K_i}{2})\, \log n+O(1) \qquad (3)$$

In Equation (3), n indicates the number of samples recorded in the dataset, k_i is number of the regression parameters in model M_i and O (1) is the error term.

In addition to the computation posterior probability of the models that enter in BMA analysis, the posterior probability can be obtained for each of variables (genes) that enter in the analysis.

This information is useful in facilitating biological discussion and it helps to identify suitable predictor variables.

Let the ($b_i\neq0$) shows that the regression parameter for gene X_i exists in the vector of regression parametersθ_i. In this case, posterior probability which gene X_i is a suitable predictor variable, is written as bellow:

$$pr(b_i\neq0|TD) = \Sigma_{M_S \text{ where gene is relevant}}\, pr(M_s|TD) \qquad (4)$$

Posterior probability of gene X_i is sum Posterior probabilities of all models in subset S that contain gene X_i (Annest et al., 2009).

Use of Iterative BMA method is unsuitable for microarray data. Because microarray dataset contains thousands of genes, while the implementation of leaps and bounds algorithm is slow when there are more than 45 variables (Annest et al., 2009). Performing a preprocessing step is proposed for dimension reduction and ranking genes (Annest et al., 2009). In present study, the univariate Cox proportional hazard model was used in preprocessing step. With parameters estimation of univariate Cox proportional hazard model, genes are sorted on the basis of their log likelihood in descending order. Then ranked genes from 1 to 1000 are selected that were more associated with survival. For selecting model, we fitted all possible models with at most m variables for the ranked genes from 1 to m (m≤25). We selected 50 (nbest=50) models with the best fit among them on the basis of leaps and bounds algorithm. For reducing set of selected models, we used the Occam's window algorithm and we removed the models with a posterior probability of less than the threshold (less than 5% posterior probability the strongest model). Then, we calculated the posterior probability of m genes by using the selected models and we eliminated genes with the low posterior probability. We replaced deleted genes with next genes in the terms of rank. These steps continually repeated until through 1000 genes processed. Finally, the top models are obtained with maximum 25 genes in each model.

To evaluate the performance of iterative BMA method, risk scores of test group separated into two risk groups. The overall risk score for a patient is considered weighted average of the risk scores calculated for each model Mi in the set S. The equation is as bellow:

$$\sum_{i\in s}(x_j^V\ \hat{\theta}_i)\ \mathrm{pr}(M_i|TD) \qquad (5)$$

Note that maximum likelihood estimation of the predictive coefficients and the posterior probability of the models is obtained from training dataset and the expression score of each gene into model for a patient (x_j^V) achieved of test dataset. Finally, we determined cutpoint

Table 1. Selected Genes by the Iterative BMA, Their Corresponding Posterior Probabilities, Log Likelihood Ranking and Description of Genes

No.	Selected genes	Posterior probability	Univariate Cox ranking	Gene description
1	PSMA7	0.945	20	|AF022815||Hs.233952|Proteasome (prosome, macropain) subunit, alpha type, 7
2	ALDOB	1	88	|X02747|*H72098|Hs.530274|Aldolase B, fructose-bisphosphate
3	GLIPR1	1	213	|X91911|*AA807145|Hs.205558|GLI pathogenesis-related 1 (glioma)
4	ANPEP	0.806	357	|M22324|*T73440|Hs.1239|Alanyl (membrane) aminopeptidase (aminopeptidase N, aminopeptidase M, micros
5	STAT4	1	473	|L78440|*H42789|Hs.80642|Signal transducer and activator of transcription 4
6	IGJ	0.973	488	|NM_144646|*AA714365|Hs.381568|Immunoglobulin J polypeptide, linker protein for immunoglobulin alpha
7	TCF12	0.519	501	|M65209||Hs.511504|Transcription factor 12 (HTF4, helix-loop-helix transcription factors 4)
8	MAP2K3	0.858	834	|D87116|*H08749|Hs.514012|Mitogen-activated protein kinase kinase 3
9	UBE2A	0.021	883	|M74524|*AA804394|Hs.379466|Ubiquitin-conjugating enzyme E2A (RAD6 homolog)
10	Ubiquitin-conjugating enzyme	0.643	925	|AI436620|*AI436620|Hs.385986|Ubiquitin-conjugating enzyme E2B (RAD6 homolog)
11	MAPK3	0.367	937	|X60188|*AA826939|Hs.861|Mitogen-activated protein kinase 3
12	HLA-E	0.019	940	|X56841|~H42063|Hs.381008|Major histocompatibility complex, class I, E
13	MGC27165	0.373	942	|BX640847|~H45437|Hs.497723|Hypothetical protein MGC27165
14	Transcribed locus	0.406	950	|AI440068|AI440068|Hs.165153|Transcribed locus
15	TRIM26	0.039	962	|BC021115|~AA421953|Hs.485041|Tripartite motif-containing 26
16	CD63	0.04	973	|NM_001780|*AA430369|Hs.445570|CD63 antigen (melanoma 1 antigen)
17	MGC61571	0.084	983	|BX648671||Hs.437336|Hypothetical protein MGC61571
18	KIAA0033	0.985	986	|D26067||Hs.501865|KIAA0033 protein
19	IGLL1	0.049	989	|M27749|*W73790|Hs.348935|Immunoglobulin lambda-like polypeptide 1
20	V01555	0.901	990	|V01555|||
21	Ubiquitin-conjugating enzyme	0.181	992	|NM_016336|*AA488853|Hs.163776|Ubiquitin-conjugating enzyme E2, J1 (UBC6 homolog
22	M33374	0.964	996	|M33374|*H46693||
23	CLAM2	0.364	997	|D45887|*AA761097|Hs.468442|Calmodulin 2 (phosphorylase kinase, delta)
24	ADORA2A	0.008	999	|M97370||Hs.197029|Adenosine A2a receptor
25	HOXD9	0	1000	|NM_014213||Hs.236646|Homeo box D9

to identify high-risk group from low-risk group by using the risk score of patients in the train group. In present study, cutpoint acquired 60%. In other words, low-risk group contains 60% of the low boundary of the risk scores and high-risk group contains 40% of the high boundary of the risk scores.

Also, predictive efficiency of two groups of patients has been assessed by Kaplan-Meier survival curve and log- rank test. For data analysis in confidence level 95%, the bioconductor package iterativeBM in R12.2.0 environment has been used. This package is available at http://www.bioconductor.org/.

Results

In this study, we performed iterative Bayesian model averaging method for dataset of the Mantle Cell Lymphoma patients. The 132 Competitive models were selected based on the posterior probabilities and the cutpoint 60%. Number of variables in each model was changing from 9 to 16 and the average number of variables was 12.68 genes per model. The posterior probabilities and univariate log likelihood ranking related to the 25 selected genes have been shown in Table 1. Among the 25 genes, HOXD9 gene with zero posterior probability has not been selected. Result of this table is showing that Among the 25 selected genes by BMA algorithm, 7 genes (28%) have rank above 502. Except 7 mention genes, the others selected genes have poor univariate ranks, so that, the highest ranking obtained from Cox proportional hazard model among them is 834 through 1000. In addition, the average ranking of the three genes with posterior probability 1 was 258.

The test group was consisted 31 patients. Iterative BMA method assigned 13 patients to high-risk category and 18 patients to low-risk category. Among 31 patients, 9 patients had been censored. One patient placed in high-risk group, from the 9 censored patients, While 12 patients of the uncensored group assigned to high-risk group. Table 2 shows the number of patients in each group.

To evaluate results, the Kaplan-Meier survival curves and the log- rank test were used. Figure 1 shows the Kaplan-Meier survival analysis curve. The based on this figure, there is consistently difference between the survival curves of two groups which survival length of people of high-risk group is very lower than people of low-risk group. Also, for the test sample, based on log-rank test, value of chi-Square obtained 21.1 (P<0.001) that indicates the strength of iterative BMA method for

Table 2. Censored and Uncensored Samples, in Each Risk Group

Group	Censored	Uncensored	Total
High risk	1	12	13
Low risk	8	10	18
Total	9	22	31

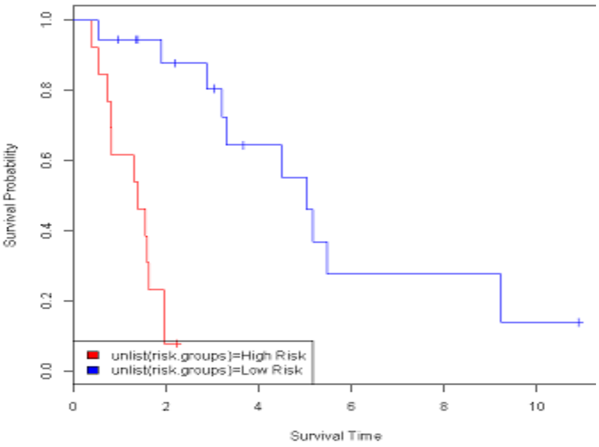

Figure 1. Kaplan-Meier Survival Analysis Curve

Table 3. The Cox Model Estimated Parameters and P-values

No.	Variable (gene)	β	SE	Wald	df	p-value	e(β)
1	PSMA7	2.31	0.9	6.62	1	0.01	10.02
2	ALDOB	-1.74	0.38	20.65	1	0	0.18
3	GLIPR1	-2.58	0.58	20.04	1	0	0.08
4	ANPEP	1.5	0.76	3.85	1	0.05	4.47
5	STAT4	-1.56	0.39	16.03	1	0	0.21
6	IGJ	0.78	0.22	12.75	1	0	2.18
7	TCF12	1.74	0.75	5.36	1	0.02	5.72
8	MAP2K3	-1.8	0.5	12.83	1	0	0.17
9	Ubiquitin-conjugating enzyme	-1.58	0.55	8.34	1	0	0.21
10	MAPK3	2.09	0.8	6.8	1	0.01	8.05
11	Transcribed locus	-0.87	0.43	4.17	1	0.04	0.42
12	TRIM26	1.2	0.61	3.81	1	0.05	3.32
13	KIAA0033	-1.48	0.47	9.81	1	0	0.23
14	V01555	0.41	0.13	9.33	1	0	1.51
15	M33374	-2.02	0.74	7.39	1	0.01	0.13
16	CLAM2	1.52	0.69	4.9	1	0.03	4.57

distinguishing the two risk groups from each other. For 25 genes simultaneously, the based on Cox proportional hazard model, 16 genes were significant (P<0.001). Table 3 shows estimation of the Cox proportional hazard model parameters and the level of significant for 16 genes.

Discussion

In present study, the iterative BMA method has been applied for using in survival analysis with high dimensional microarray data. Using this method numbers of 8810 genes have been reduced to 25 genes and 132 top survival models have been selected. Selected models were rather simple and including 9 to 16 genes. Also, Posterior probabilities of the selected genes were calculated by the iterative BMA. Values of the posterior probability of chosen genes was indicating overall contribution that gene into the patient risk score across all selected models. Finally, patients were separated into two risk groups: low-risk group and high-risk group, with very high significantly. The results of this study showed that iterative BMA method is able to separate risk groups with very high significantly. The Kaplan-Meier survival curves and the log-rank test implied the high power of iterative BMA method to predictive survival.

Algorithm BMA, in preprocessing step, Cox proportional hazard model fits for each genes and genes orders based on their log likelihood in descending order. Therefore, there are among selected variables comparison possibility of log likelihood values of top genes. The gene U19769 (not inclusion in selected genes) has first rank with log likelihood equal -214.57. The log likelihood for genes PSMA7 with ranking of 20 and HOXD9 with ranking of 1000, are respectively -218.41 and -288.29. So, genes with the poor ranking of log likelihood of univariate Cox often are selected by the iterative BMA algorithm and in the terms of goodness of fit are comparable with the top ranking genes. This is because that log likelihood scope of univariate Cox proportional hazard model is not extensive across the top 1000 genes. Thus, it is not surprising that selected genes by BMA that in the terms of log likelihood have poor ranking, achieve substantial predictive power when included in combinations. Previous studies have reported similar results. Hu et al, showed that gene PSMA7 is expressing in high level in the colorectal cancer sites and lymph node and liver metastatic sites while gene expression is not high in the normal colorectal tissue (Hu et al., 2008). Wan et al. (2004), Liu et al. (2007), Midorikawa et al. (2002), Li et al. (2004) respectively report that expression of genes M33374, ALDOB, KIAA0033 and V01555 are associated with Hepatocellular Carcinoma. Also, Muller et al, showed that GLIPR1 associates with prostate cancer (Muller et al., 2010). Chang et al. (2009) in their study, reported role of STAT4-deficient in the impaired development of human Th1 cells for posttransplantation patients (Chang et al., 2009). Park et al. (2008) detected existence the aberrant methylation for gene MAP2K3 in at least one lung adenocarcinoma cell lines (Park et al., 2008). Ma et al. (2009) report connection IGJ gene with breast cancer (Ma et al., 2009).

Although, the achieved results in this study were hopeful, the iterative BMA method has some limitations and can be extended further. Such as, determining the optimum number of risk groups and proposing statistical methods, for the assessment of different calculation methods. It is suggested that validation the chosen genes obtained of the iterativeBMA method collaborate with genetic and clinical studies.

Iterative BMA algorithm has the high accurate and strength for survival analysis. This method is able to identify a few numbers of predictive variables among many variables (genes) in a microarray dataset. So, it can be used as a low-cost diagnostic tool in clinical researches. This multivariate technique computes the model uncertainty through the averaging over the posterior probabilities of the strongest competitive models. Multivariate feature iterative BMA with the ability to calculate the model uncertainty makes this method as an interesting pattern for extraction predictive genes of high-dimensional biological data.

References

Adamson P, Bray F, Costantini A, et al (2007). Time trends in the registration of Hodgkin and non- Hodgkin lymphomas in Europe. *Eur J Cancer*, **43**, 391-401.

Annest A, Bumgarner RE, Raftery AE, et al (2009). Iterative Bayesian Model Averaging: a method for the application of survival analysis to high-dimensional microarray data. *BMC Bioinformatics*, **10**, 72.

Campo E, Raffeld M, Jaffe ES (1999). Mantle-cell lymphoma. *Seminars Hematol*, **36**, 115-27.

Chang H-C, Han L, Goswami R, et al (2009). Impaired development of human Th1 cells in patients with deficient expression of STAT4.

Fisher p (2005). Non-Hodgkin's lymphoma. *Practice Oncol*, 1948-52.

Furnival G, Wilson R (1974). Regression by Leaps and Bounds. *Technometrics*, **16**, 499-511.

Hu X-T, Chen W, Wang D, et al (2008). The proteasome subunit PSMA7 located on the 20q13 amplicon is overexpressed and associated with liver metastasis in colorectal cancer. *Oncology Reports*, **19**, 441-6.

Kass RE, Raftery AE (1993). Bayes factors and model uncertainty.

Li J, Duan Y, Ruan X (2007). A novel hybrid approach to selecting marker genes for cancer classification using gene expression data. *Bioinformatics Biomedical Engineering*, 264-7.

Li W, Wu BA, Zeng YM, et al (2004). Epstein-Barr virus in hepatocellular carcinogenesis. *World J Gastroenterol*, **10**, 3409-13.

Liu Y, Zhu X, Zhu J, et al (2007). Identification of differential expression of genes in hepatocellular carcinoma by suppression subtractive hybridization combined cDNA microarray. *Oncol Reports*, **18**, 943-51.

Ma X-J, Dahiya S, Richardson E, et al (2009). Gene expression profiling of the tumor microenvironment during breast cancer progression. *Breast Cancer Res*, **11**.

Madigan D, Raftery AE (1994). Model selection and accounting for model uncertainty in graphical models using occamis Window. *J American Statistical Associat*, **89**, 1335-46.

Midorikawa Y, Tsutsumi S, Taniguchi H, et al (2002). Identification of genes associated with dedifferentiation of hepatocellular carcinoma with expression profiling analysis

Jpn J Cancer Res, **93**, 636-43.

Muller I, Wischnewski F, Pantel K, et al (2010). RPerseoarmch aortitclee r- and cell-specific epigenetic regulation of CD44, Cyclin D2, GLIPR1 and PTEN by Methyl-CpG binding proteins and histone modifications. *BMC Cancer*, **10**, 1-15.

Park JC, Chae YK, Son CH, et al (2008). Epigenetic silencing of human T (brachyury homologue) gene in non-small-cell lung cancer. *Biochemical Biophysical Res Communicat*, **365**, 221-6.

Raftery A (1995). Bayesian model selection in Social Research. *Sociological methodol*, **25**, 111-96.

Rosenwald A, Wright G, Wiestner A, et al (2003). The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell*, **3**, 185-97.

Rust R, Visser L, van J, et al (2005). High expression of calcium-binding proteins, S100A10, S100A11 and CALM2 in anaplastic large cell lymphoma. *British J Haematol*, **131**, 596-608.

Swerdlow SH, Williams ME (2002). From centrocytic to mantle cell. lymphoma: a clinicopathologic and molecular review of 3 decades. Human Pathology, 33, 7-20.

Wan D, He M, Wang J, et al (2004). Two variants of the human hepatocellular carcinoma-associated HCAP1 gene and their effect on the growth of the human liver cancer cell line Hep3B. *Genes, Chromosomes Cancer*, **39**, 48-58.