FULL LENGTH	Iranian Biomedical Journal x (x): x-x x x

Application of Sparse Linear Discriminant Analysis and Elastic Net for Diagnosis of IgA Nephropathy: Statistical and Biological Viewpoints

Tahereh Mohammadi Majd[1], Shiva Kalantari[*2], Hadi Raeisi Shahraki[3], Mohsen Nafar[4], Afshin Almasi[5], Shiva Samavat[6], Mahmoud Parvin[7] and Amirhossein Hashemian[8,9]

[1]Department of Biostatistics and Epidemiology, Kermanshah University of Medical Sciences, School of Public Health, Kermanshah, Iran; [2]Chronic Kidney Disease Research Center, Labbafinejad Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran; [3]Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran; [4]Urology-Nephrology Research Center, Labbafinejad Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran; [5]Department of Biostatistics and Epidemiology, School of Public Health, Kermanshah University of Medical Sciences, Kermanshah, Iran; [6]Department of Nephrology, Labbafinejad Medical Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran; [7]Department of pathology, Labbafinejad Medical Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran; [8]Research Center for Environmental Determinants of Health (RCEDH), Kermanshah University of Medical Sciences, Kermanshah, Iran; [9]Department of Biostatistics and Epidemiology, Faculty of Health, Kermanshah University of Medical Sciences, Kermanshah, Iran

Received 9 October 2017; revised 31 December 2017; accepted 3 January 2017

ABSTRACT

Background: IgA nephropathy (IgAN) is the most common primary glomerulonephritis diagnosed based on renal biopsy. Mesangial IgA deposits along with the proliferation of mesangial cells are the histologic hallmark of IgAN. Non-invasive diagnostic tools may help to prompt diagnosis and therapy. The discovery of potential and reliable urinary biomarkers for diagnosis of IgAN depends on applying robust and suitable models. Applying two multivariate modeling methods on a urine proteomic dataset obtained from IgAN patients, and comparison of the results of these methods were the purpose of this study. **Methods**: Two models were constructed for urinary protein profiles of 13 patients and 8 healthy individuals, based on sparse linear discriminant analysis (SLDA) and elastic net regression methods. A panel of selected biomarkers with the best coefficients were proposed and further analyzed for biological relevance using functional annotation and pathway analysis. **Results**: Transferrin, α1-antitrypsin, and albumin fragments were the most important up-regulated biomarkers, while fibulin-5, YIP1 family member 3, prasoposin, and osteopontin were the most important down-regulated biomarkers. Pathway analysis revealed that complement and coagulation cascades and extracellular matrix-receptor interaction pathways impaired in the pathogenesis of IgAN. **Conclusion**: SLDA and elastic net had an equal importance for diagnosis of IgAN and were useful methods for exploring and processing proteomic data. In addition, the suggested biomarkers are reliable candidates for further validation to non-invasive diagnose of IgAN based on urine examination.

Keywords: IgA nephropathy, Proteomics, Biomarker, Diagnosis

Corresponding Author: Shiva Kalantari
Chronic Kidney Disease Research Center, Labbafinejad Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran;
Tel.: (+98-21) 22594197; Fax: (+98-21) 22580201; E-mail: shiva.kalantari@sbmu.ac.ir

INTRODUCTION

IgA nephropathy (IgAN) is the most common primary glomerulonephritis in the world[1]. IgAN is diagnosed based on invasive renal biopsy by evidence of mesangial IgA deposits along with proliferation of mesangial cells[1,2]. The disease has a wide spectrum of clinical presentations, ranging from asymptomatic microscopic hematuria in mild stages to macroscopic hematuria with heavy proteinuria in a

more severe course that results in rapid deterioration of renal function[3]. Approximately 20-40% of patients with IgAN end up to end-stage renal disease within 20 years after diagnosis[4]. Thus, a timely and non-invasive diagnosis and a more clear insight of pathogenesis may help to save patients from kidney failure. Recently, numerous high-throughput studies have attempted to identify more applicable, reliable and non-invasive biomarkers for renal diseases[5-10], instead of the application of non-specific biochemical factors and invasive diagnostic methods. Several omics analyses have been used for identifying biomarkers for IgAN[6,11-13]; however, the reliable non-invasive diagnostic biomarkers and the impaired biological pathways involved in the process of the disease need to be determined. These high-throughput data, which are obtained via omics technologies (e.g. proteomics, transcriptomics, metabolomics, genomics, meta-genomics, epigenomics, etc.), are characterized by the dimensionality and complexity[14]. Dimensionality refers to hundreds to thousands of variables or "p" (e.g. genes, proteins, metabolites, etc.), whereas the number of samples or "n" are relatively small[15]. When these data are high-dimensional and multicollineared, the univariate analysis, which ignores the covariance between the variables and is prone to false positives, seems not to be appropriate[16]. Therefore, univariate analysis could not reflect the biological relationship between the correlated variables.

To address the extraction of biological information due to joint impacts of variables and dimension reduction, different feature selection and classification methods such as least absolute shrinkage and selection operator (LASSO)[17], adaptive LASSO[18], partial least squares discriminant analysis (PLS-DA)[9], sparse linear discriminant analysis (SLDA)[19], elastic net-type regularized regression[20], and smoothly clipped absolute deviation (SCAD)[21] have been established over the past decades. The judge on the applicability of the analysis method depends on the aim of study and quality of data. Elastic net and SLDA are two of the popular methods for high-throughput data analysis. Elastic net is a linear combination of Ridge[22] and Lasso[17] with both shrinkage and automatic biomarker selection, whereas SLDA performs discriminant analysis with penalty coefficients to obtain sparsity according to chosen variables[17,20]. We selected these two methods because elastic net has advantages of both Ridge and LASSO regression (when the penalty weight (α) is equal to 1, it produces the LASSO regression, and when $\alpha = 0$, it produces Ridge regression. Accordingly, in elastic net, penalty weight is $0 \leq \alpha \leq 1$. Furthermore, the results of analyzing high-throughput data are more interpretable than using ordinary linear discriminant analysis (LDA), and it is extensible to non-normal data. In this study, we compared the efficiency of these two methodologies to obtain reliable diagnostic biomarkers for IgAN. The results were then compared with PLS-DA in our previous paper[10]; a panel of selected biomarkers based on these three models are suggested, and biological relevance of identified biomarkers are discussed.

MATERIALS AND METHODS

Description of urine proteome dataset

In order to compare the performance of two methods (SLDA and elastic net) for diagnosis of IgAN, urine protein profile of patients with IgAN and of healthy subjects, which was obtained using nanoscale liquid chromatography-high resolution tandem mass spectrometry, was used[10]. The dataset is made up of the urinary protein profiles of 13 patients (11 men and 2 women) and 8 (6 men and 2 women) healthy volunteers as the control group with the mean age of 33 and 34.5 years of age, respectively, each with 493 variables. The samples were collected from patients who were referred to Labbafinejad Medical Center in Tehran (Iran) during 2011 to 2012. Demographic, histopathologic and clinical data of the patients are tabulated in Table 1.

Sparse linear discriminant analysis (SLDA)

LDA, which uses a linear combination of features as the criterion for classification, has often been shown to produce the best classification results for low dimensional data[23-26]. Although LDA is popular because of its simplicity and predictive ability, it fails to work when there are too many correlated predictors, and when the number of features is exceeded than sample size[19,26,27]. In this condition, which is mainly true in omics data, the result is difficult to interpret. To solve the limitations of LDA, the SLDA was defined. SLDA led to have a model with more accurate estimation and more accurate prediction capabilities as well as a higher power. Suppose that X is an $n \times p$ matrix, where p is the number of predictor variables, and n is the number of samples. Also, x_i and x_j are i^{th} row and j^{th} column of the matrix X, respectively. The number of observations in k^{th} group is shown with C_k, that is a subset of[12] $n_k = |C_k|$, $\sum_{k=1}^{n} n_k = n$. An estimate for the within-class covariance matrix $\hat{\Sigma}_W$ and an estimate for the covariance matrix between classes $\hat{\Sigma}_b$ can be defined as follows, respectively:

$$\hat{\Sigma}_W = \frac{1}{n}\sum^{K}\sum (X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^T$$

Table 1. Demographic, histopathologic, and laboratory data of patients

No.	Age (Y)	Sex	MEST score	WHO Stage	MH	EnH (%)	ExH (%)	IFTA (%)	Sclerosis (%)	Inflammation (%)	sCr (mg/dl)	Proteinuria (mg/24h)
1	28	M	M1E0S1T2	V	severe	NO	NO	80	83	25	4.6	2330
2	34	M	M1 S1 E0 T1	V	severe	NO	NO	60	73	25	2.3	2640
3	18	M	M1 S0 E0 T1	III	mild	NO	NO	20	NO	50	0.9	1000
4	45	M	M1 S1 E1 T0	III	mild	<50	NO	15	29	NO	1.3	720
5	47	M	M1 S0 E0 T0	II	mild	NO	NO	10	NO	25	1.1	6420
6	41	M	M1 S0 E0 T0	II-III	moderate	NO	NO	NO	5	NO	0.5	520
7	29	M	M1 S1 E1 T2	V	mild	<50	<50	80	50	>50	7.7	7020
8	28	F	M0 S1 E0 T0	III	mild	NO	NO	20	NO	NO	0.7	1680
9	34	M	M1 S0 E0 T0	II-III	moderate	NO	NO	5	14	NO	1	1310
10	32	M	M1 S1 E0 T1	IV	moderate	NO	<50	40	30	35	1.8	4100
11	29	M	M1 S0 E1 T2	V	severe	NO	NO	80	78	<25	5	6000
12	23	F	M1 S1 E1 T1	III	moderate	<50	NO	20	15	NO	1.2	800
13	51	M	M1 S1 E0 T0	II	moderate	NO	NO	15	NO	NO	1.7	4600

No. number of patients; MH, mesangial hypercellularity; EnH, endo-capillary hypercellularity; ExH, extra-capillary hypercellularity; IFTA, interstitial fibrosis-tubular atrophy; sCr, serum creatinine. MEST score is acronym of mesangial proliferation, endocapillary hypercellularity, sclerosis/adhesions, tubular atrophy/interstitial fibrosis.

$$\hat{\Sigma}_b = \frac{1}{n} X^T X - \hat{\Sigma}_W = \frac{1}{n} \sum_{k=1}^{K} \hat{\mu}_k \hat{\mu}_k{}^T$$

That $\hat{\mu}_k$ is the sample mean vector for k^{th} class. Finally, the following command acquires k^{th} penalized discriminant vector:

$$\text{maximize}_{\beta_k} \left\{ \beta_k{}^T \hat{\Sigma}_b{}^k \beta_k - P_k(\beta_k) \right\} \text{ subject to } \beta_k{}^T \tilde{\Sigma}_W \beta_k \leq 1,$$

Where P_k is a convex penalty function on the k^{th} discriminant vector, and $\tilde{\Sigma}_W$ is a positive definite estimate for Σ_W. SLDA analysis was performed using PenalizedLDA package.

Elastic net regression analysis

The elastic net-type regularized regression (e.g., ridge[22], lasso[17], elastic net[20], etc.) is a popular data analysis method for identifying features based on high, dimensional omics dataset[28]. For less variability and construction a reliable model in case of datasets with correlated features, elastic net regression reduce the variance of the model by constraining the size of the regression coefficients[29]. This method of feature selection is called regularization based on imposed constraint (i.e. penalty) on the coefficients. The elastic net regression analysis acts as a bridge between the ridge and lasso regressions. The L_2-norm (the sum of squared values), and the L_1-norm (the sum of absolute values) of the coefficients are considered as penalty in ridge and lasso regressions, respectively, while a hybrid of these two penalties is considered in elastic net regression.

$$L(\alpha, \beta) = |Y - \beta X|^2 + \lambda(\alpha|\beta|_1 + (1-\alpha)\beta^2)$$

Where $|\beta|_1$ is L_1-norm of the vector of regression coefficients, and $|Y - \beta X|^2$ is the sum of the squared residuals from the fit[29]. The value of the λ and α parameters can be estimated by performing cross-validation method. For n observation, if $\beta^T = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of p variables, elastic net logistic regression is defined as follow:

$$L(\beta; \lambda) = l_n(\beta) + \lambda \sum_{j=1}^{p} \left[(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right]$$

Where ln (β) is a maximum likelihood estimator for logistic regression, α is a value between zero and one, and λ is a positive constant called tuning parameter that manages the shrinkage degree[29]. Elastic net regression was performed using package glmnet in R 3.3.1 software. Parameters for evaluating the agreement of the models, such as cure agreement and kappa, were calculated, and Bland-Altman plot was drown. The area under curve (AUC) values was calculated using receiver operating characteristic (ROC) curves for evaluation the accuracy of the models. In order to estimate the optimum amounts of shrinkage in both elastic net and SLDA methods, fivefold cross validation was used. In addition, standard error of coefficients was obtained using 500 times replication of bootstrap methods. Bland-Altman plot was drawn in MedCalc 14.0 software, and the other calculations were performed using R 3.3.1 software.

Table 2. Summary of models

Models	Zero	None Zero	Total
SLDA	360	133	493
EN	373	120	493

Functional analysis of identified biomarkers

To further understand the biological relevance of the characterized biomarkers, we performed gene ontology analysis using Cytoscape v 3.4.0 software and a Cytoscape plug-in named ClueGO (version 2.2.5)[30,31]. ClueGO is widely used for analysis and visualization of functionally related genes. Gene ontology analysis composed of three terms, including "biological process", "cellular component", and "molecular function" was performed, and the enriched pathways in Kyoto Encyclopedia of Genes and Genomes database (http://www.genome.jp/kegg/) were also identified. The statistical test used for the enrichment was based on a two-sided hypergeometric option with a Benjamini-Hochberg correction, a *p* value less than 0.05, and a kappa score threshold of 0.4. The minimum number of genes was considered 3.

RESULTS

Biomarker identification based on elastic net and SLDA models

In this study, we examined the effect of 493 variables in urinary protein profile of IgAN patients and healthy subjects. Univariate analysis using Mann-Whitney test revealed that there was a significant difference ($p < 0.05$) between the case and control groups in 144 out of 493 variables (the results not shown). Because the sample size was small, we directly used fivefold cross-validation to determine the training data and the test data and selected the best parameters (e.g. λ and α) for the methods. For assessing simultaneous effects of aforementioned variables on IgAN disease, elastic net and SLDA models were fitted based on $\lambda = 0.005$ and $\lambda = 0.06$, respectively. The results of two models indicated that 133 out of 493 variables were effective in discrimination of IgAN in SLDA model, whereas 120 predictive variables were important in elastic net model. Summary of models are shown in Table 2. In this Table, 36 and 37 most important variables in terms of the highest coefficient were reported as discriminative diagnostic biomarkers between two groups for elastic net and SLDA models, respectively. The coefficients of elastic net regression and SLDA for the most effective variables in bootstrap method are shown in Figure 1. There was a good agreement between two models since 30 of selected biomarkers were identical (Table 3), and cure agreement and kappa were 90% and 75%, respectively.

Bland-Altman plot was drawn based on the rank of the importance variables for emphasizing the agreement between the models (Fig. 2). Table 4 describes the variables that were different between the models. ROC curve revealed that the AUC for both models were 100%, and no misclassification was observed. Also, the optimum cutoff point for probability of IgAN was obtained as 0.52 in elastic net regression using ROC curve analysis. We considered the protein IDs from the Table 2 as the most important diagnostic biomarkers, because these biomarkers were remained significant with high coefficient value in both models. The top four up- and down-regulated biomarkers are tabulated in Table 5 with calculated sensitivity and specificity.

Functional annotation of IgAN-related biomarkers

To better understand the biological functions of the most important discriminative proteins, we carried out functional enrichment analyses via ClueGO. The integrative list of the biomarkers identified by two models composed of 43 proteins (Tables 2 and 3) was analyzed by GO terms and pathways. Only GO terms with a corrected *p* value < 0.05 were considered statistically significant. Three major groups, including acute-phase response ($p = 24 \times 10^{-6}$), fibrinolysis ($p = 35.0 \times 10^{-6}$), and platelet degranulation ($p = 3.1 \times 10^{-9}$), encompassing seven terms of biological process were remained significant. The significant terms and their nodes are displayed in Figure 3A. As shown in Figure 3B, basement membrane ($p = 2.1 \times 10^{-6}$), secretory granule lumen ($p = 15 \times 10^{-9}$), and blood microparticle ($p = 250 \times 10^{-12}$) were the important biomarkers enriched in three clusters composed of seven terms of cellular component. The GO levels were different for each term, and vary between 2 to 12. However, each term was reported under multiple levels from general nodes (higher parents) to more specific child nodes (lower nodes). In contrast, no GO term was enriched for the categories of molecular function. The results of pathway enrichment analysis revealed two significant pathways: complement and coagulation cascades ($p = 1.9 \times 10^{-5}$) and extracellular matrix (ECM)-receptor interaction ($p = 1.9 \times 10^{-5}$). The enriched pathways and their nodes are displayed in Figure 4.

DISCUSSION

IgAN is the most common type of primary glomerulonephritis worldwide. This disease has a significant morbidity and leads to end-stage renal disease in about 40% of patients within 20 years of diagnosis[32]. The histopathologic hallmark of IgAN

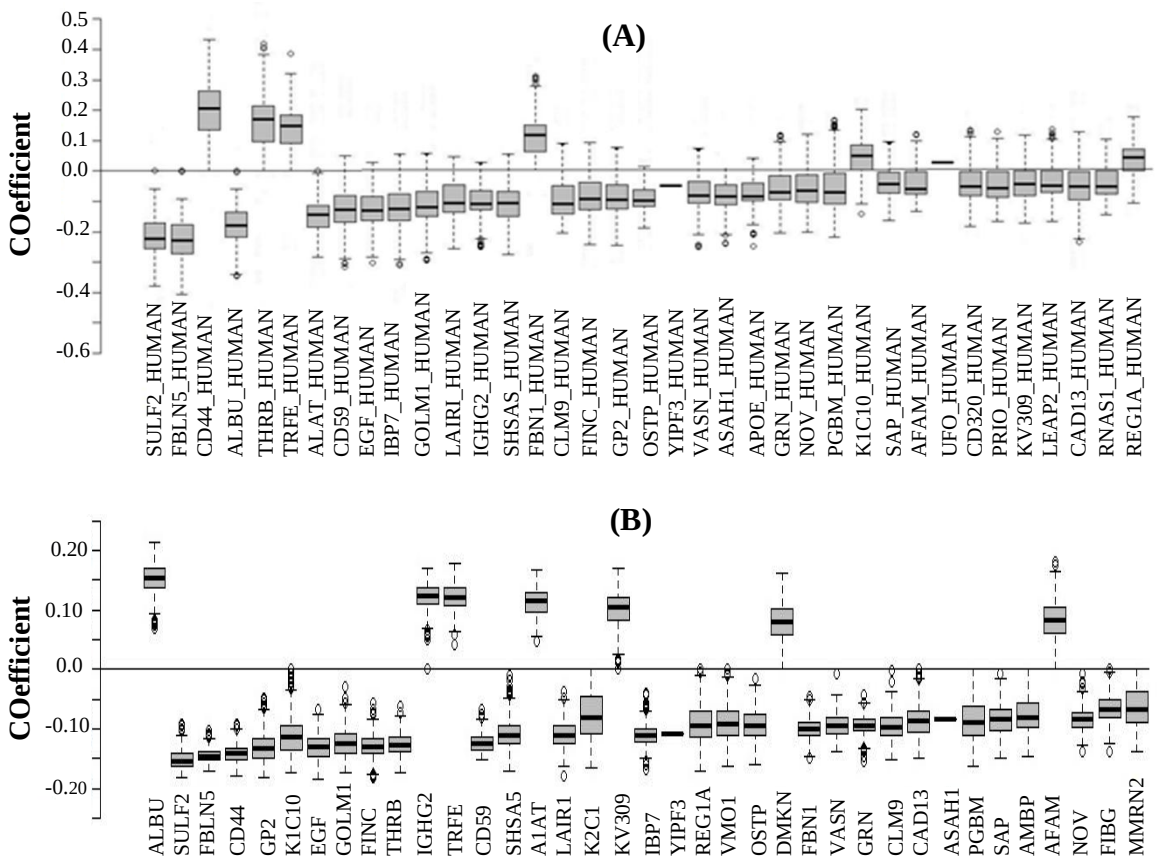

Fig. 1. The coefficients of (A) elastic net and (B) SLDA extracted in 500 times bootstrap method.

is the dominant or co-dominant deposition of IgA in the glomerular mesangium that is usually accompanied by mesangial cellular proliferation and expansion of the ECM[33]. Although renal biopsy involves a risk of morbidity due to bleeding complications, it has currently been considered the only reliable diagnostic approach for IgAN as other glomerular diseases[12]. Nonetheless, other alternative non-invasive approaches appear to be necessary for reducing the difficulties of biopsy and improving its reliability. The ideal non-invasive method is analyzing urine specimen using omics technologies such as urinary proteomics and metabolomics. However, analyzing high-dimensional data producing by these techniques are challenging and require appropriate multivariate modeling. Two popular multivariate models that are compatible with the nature of high-dimensional data are SLDA and elastic net. We have compared the application of these two robust models to identify diagnostic protein biomarkers for IgAN. The good reproducibility of the results from two models indicates the efficiency and the power of penalty-based regression models for biomarker discovery researches. Since the accuracy of both models was 100%, there was no superiority between them in this case. Therefore, we highlighted the importance of a subset of 30 shared biomarkers obtained using two models with highest regression coefficients and less error (Table 3). The top four

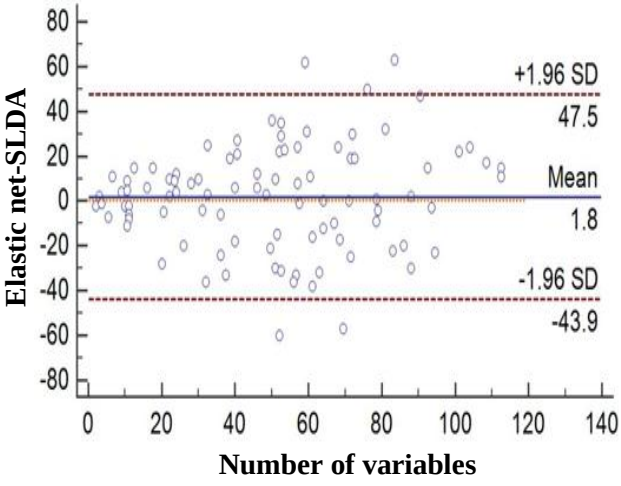

Fig. 2. Bland-Altman plot for SLDA and elastic net models.

Table 3. List of identical biomarkers that were significant in both models

Protein ID	SLDA		Elastic Net		FC	Direction
	Coefficient	SE	Coefficient	SE		
TRFE	0.21	0.16	0.14	0.02	7.8	↑
A1AT	0.19	0.14	0.13	0.02	7.7	↑
ALBU	0.24	0.20	0.19	0.02	4.3	↑
AFAM	0.08	0.07	0.10	0.03	3.7	↑
IGHG2	0.14	0.10	0.15	0.02	2.9	↑
KV309	0.08	0.06	0.13	0.03	2.6	↑
FBLN5	-0.26	0.19	-0.17	0.01	11.8	↓
YIPF3	-0.13	0.10	-0.12	0.02	10.6	↓
SAP	-0.08	0.07	-0.10	0.03	10.3	↓
OSTP	-0.13	0.09	-0.11	0.02	9.4	↓
SULF2	-0.30	0.28	-0.18	0.02	9.0	↓
CD44	-0.26	0.21	-0.17	0.02	8.9	↓
CD59	-0.17	0.14	-0.14	0.02	7.5	↓
K1C10	-0.09	0.08	-0.16	0.03	4.7	↓
FBN1	-0.14	0.11	-0.11	0.02	4.2	↓
IBP7	-0.16	0.11	-0.13	0.02	4.2	↓
VASN	-0.12	0.09	-0.11	0.02	3.6	↓
LAIR1	-0.16	0.11	-0.13	0.02	3.6	↓
GP2	-0.13	0.11	-0.17	0.02	3.3	↓
REG1A	-0.08	0.06	-0.12	0.03	3.1	↓
SHSA5	-0.14	0.10	-0.13	0.02	3.1	↓
CLM9	-0.14	0.11	-0.11	0.02	2.7	↓
FINC	-0.14	0.10	-0.16	0.02	2.6	↓
PGBM	-0.09	0.07	-0.11	0.03	2.5	↓
GRN	-0.10	0.08	-0.11	0.02	2.4	↓
CAD13	-0.08	0.06	-0.11	0.03	2.2	↓
EGF	-0.17	0.13	-0.16	0.02	2.2	↓
ASAH1	-0.11	0.09	-0.11	0.02	2.1	↓
GOLM1	-0.16	0.12	-0.16	0.02	1.9	↓
NOV	-0.10	0.06	-0.10	0.02	1.9	↓

SE, standard error; FC, fold change; ↑, up-regulation; ↓, down-regulation

Table 4. List of significant biomarkers that were different between two models

SLDA					Elastic Net				
Protein ID	Coefficient	SE	FC	Direction	Protein ID	Coefficient	SE	FC	Direction
THRB	-0.22	0.16	4.2	↓	K2C1	-0.13	0.04	1.1	↓
APOE	-0.11	0.08	2.5	↓	VMO1	-0.12	0.03	3.8	↓
UFO	-0.08	0.05	4.7	↓	DMKN	0.11	0.03	1.9	↑
CD320	-0.08	0.06	2.0	↓	AMBP	-0.10	0.03	2.5	↓
PRIO	-0.08	0.07	2.4	↓	FIBG	-0.10	0.02	1.9	↓
LEAP2	-0.08	0.07	4.1	↓	MMRN2	-0.09	0.03	1.9	↓
RNAS1	-0.08	0.10	3.0	↓					

SE, standard error; FC, fold change; ↑, up-regulation; ↓, down-regulation

Table 5. Panel of suggested diagnostic biomarkers for IgA nephropathy based on elastic net, SLDA, and comparison with PLS-DA

Biomarker	Sensitivity (%)	Specificity (%)
TRFE	100	100
A1AT	100	100
ALBU	100	100
AFAM	100	92.31
FBLN5	100	100
YIPF3	100	100
SAP	75	100
OSTP	100	92.31
GP2	87.5	100
CLM9	100	92.31
VASN	87.5	100
CD44	100	100
EGF	100	100
FBN1	100	100

overrepresented and underrepresented biomarkers in this list were as follows, respectively: transferrin (TRFE), α1-antitrypsin (A1AT), albumin (ALBU), afamin (AFAM), and fibulin-5 (FBLN5) , YIP1 family member 3 (YIPF3), prasoposin (SAP), and osteopontin (OSTP).

The urinary excretion of the overexpressed biomarkers in patients with IgAN have been reported earlier[9,34-36], and our results validate the role of these molecules as biomarkers. However, these are not specific for IgAN and have been detected in other glomerular diseases[5,37]. The amount of excretion of proteins in IgAN may be different from other glomerular disease. Further studies requires for proving this hypothesis. Among the overrepresented biomarkers, the evidence on increased urinary A1AT in IgAN is more reasonable. A1AT, also known as SERPINA1, is a serine protease inhibitor that is mainly produced by liver cells, but it is also synthesized by macrophages, neutrophils, activated lymphocytes and intestinal epithelial cells[9]. Injured renal tubular epithelial cells also can synthesize A1AT in response to tubulointerstitial damage[38]; hence, A1AT is absence in normal urine but detectable in other renal diseases[39]. The possible relevant reasons for A1AT up-regulation in IgAN are: (I) enhancement the synthesis of TRFE receptor or CD71, that is like a receptor for aberrant IgA on the mesangial cells in the kidney tissue of IgAN patients[40], (II) inhibition of thrombin (coagulation factor II) and induction of fibrinolysis during blood coagulation by A1AT in response to hematuria, which is typically presents in IgAN[31,41], and (III) inactivation of proteinases secreted from inflammatory cells, such as neutrophils, by A1AT in IgAN patients[42] . In addition, our finding is in agreement with Neprasova *et al.*[11] and Prikryl *et al.*[42] resultsthat had detected A1AT in the urine of patients with IgAN using labeling proteomic techniques.

There is less evidence on the biomarker role of proteins that suggested as down-regulated candidates

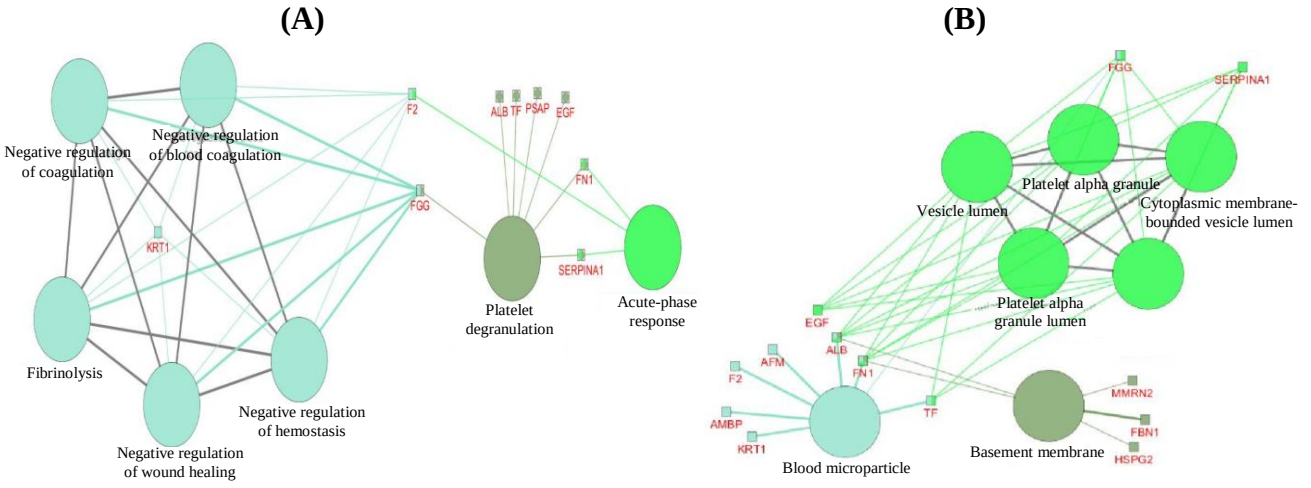

Fig. 3. The proteins encompassed by enriched biological processes (A) and cellular component (B), using Cytoscape v 3.4.0 software. The large circles represent biological processes (A) and cellular component (B), and the small rectangles represent the proteins. The circles with the same colors have the same level of significance; therefore, they are in the same GO group. In A, the blue, green, and gray circles show $p = 35.0 \times 10^{-6}$, $p = 24 \times 10^{-6}$, $p = 3.1 \times 10^{-9}$, respectively. In B, the green circles represent p value = 15×10^{-9}. The blue circle represents p value = 250×10^{-12}, and the grey circle represents p value = 2.1×10^{-6}.

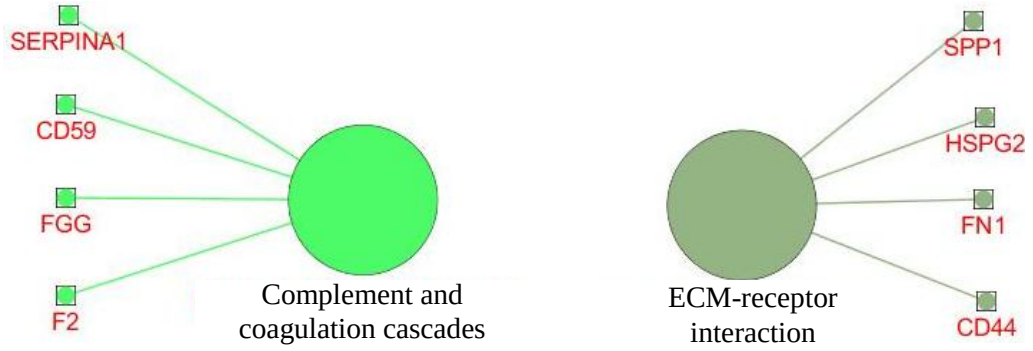

Fig. 4. Enriched pathways involved in pathogenesis of IgAN. The large circles represent pathways, and the small rectangles represent the proteins.

for IgAN unless OSTP, and they are presented in this paper for the first time. FBLN5 that is named as FBLN5 is associated with elastic fiber formation as a Ca^2-dependent elastin-binding protein and presents in interstitial renal tissue, as well as distal renal tubules[43,44]. It has a role in tissue repair, the inhibition of endogenous angiogenesis, and remodeling in response to oxidative stress-mediated renal damage[43,44]. Down-regulation of FBLN5 indicates the low capacity of repair of the injuries mediated by oxidative stress in IgAN.

SAP is a highly conserved lysosomal glycoprotein hydrolases involved in the hydrolysis of sphingolipids[45]. Accordingly, it can be postulated that the hyperlipidemia associated with IgAN with nephrotic range proteinuria might be occurred because of the decreased level of this protein in renal tissue of IgAN patients. In addition, the impaired lysosomal pathway that was reported in IgAN based on urine proteomic data[36] rationalizes the contribution of SAP as a lysosomal enzyme in the pathogenesis of IgAN.

Significant reduction of urinary excretion of OSTP in our analysis corresponds with a previous study[46], which highlight its potential role as a IgAN biomarker. A possible explanation for underrepresentation of OSTP in IgAN could be the cleavage of this protein by serine proteases that their activities increase in IgAN[47]. Furthermore, there is a strong correlation

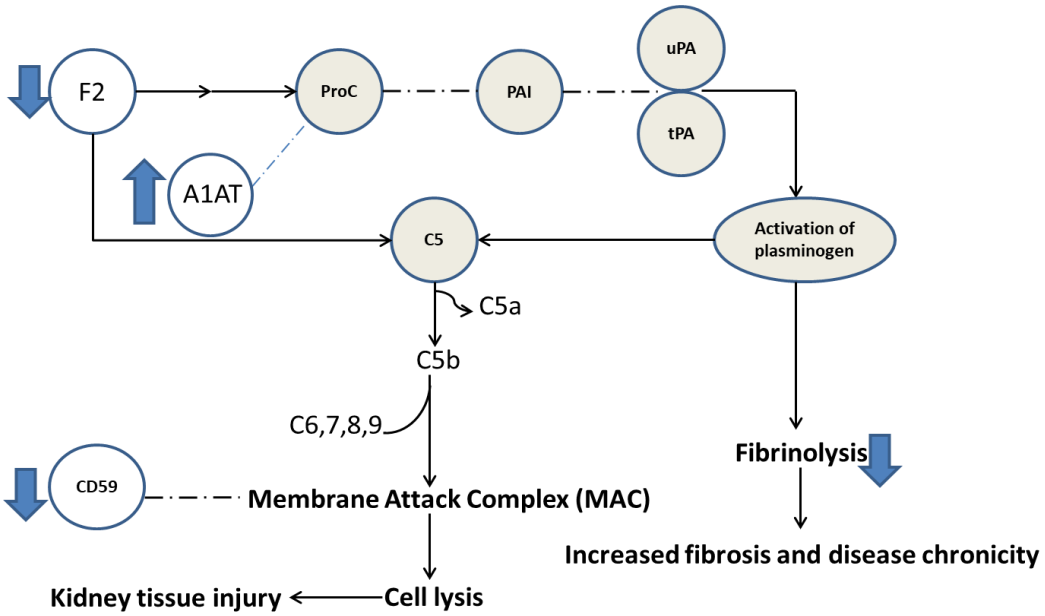

Fig. 5. The hypothetical mechanism of renal damage in IgAN mediated by coagulation and complement pathway. The open circles with their direction of changes shown by arrows represent the significant proteins that were detected in our dataset. The solid circles represent the mediators that were not in our dataset; nevertheless, connection of the detected biomarkers (in the open circles) to impaired pathways resulted in kidney injury based on literature. ProC, activated protein C; PAI, plasminogen activator inhibitor; A1AT, α1-antitrypsin; uPA, urokinase-type plasminogen activator; tPA, tissue-type plasminogen activator; F2, prothrombin; C5a, complement factor 5a; C5b, complement factor 5b; C6,7,8, and 9, complement factors 6,7,8, and 9, respectively.

between the excretion of OSTP and its receptor CD44(OSTP receptor) in our dataset (r = 0.93) that has a known contribution in IgAN[48]. CD44 is one of the seven biomarkers (i.e. IGHG2, CD44, VASN, GP2, SHSA5, CLM9, and EGF) that were remained significant in three-type analysis in the present and our previous paper[10], using SLDA, elastic net, and PLS-DA. These sets of biomarkers have been validated statistically and are important targets for experimental validation in the future studies. The contribution of and changes in the urinary level of YIPF3, a natural killer specific antigen, in patients with IgAN is reported here for the first time, and its role in pathogenesis of IgAN needs complementary experiments.

Investigation of the gene ontology and pathways that enriched in the list of significant biomarkers paved the way to translate invisible information into interpretable biological information. Involvement of acute phase response and processes related to coagulation in IgAN that have been enriched in the dataset is similar to previous studies[49-51]. The role of coagulation pathway in pathogenesis of renal diseases, its crosstalk with complement pathway, and its downstream effects on glycocalyx and formation of ECM are noteworthy nowadays[52]. Therefore, the enrichment of "complement and coagulation cascades" and ECM-receptor interaction pathways is highly relevant to the pathogenic process of IgAN. The important encompassing nodes of these pathways have been described in Figure 4. The possible mechanism of fibrosis and renal damage in IgAN based on our findings might be as follows: down-regulation of F2 in association of up-regulation of A1AT leads to a decrees in the production of activated protein C, which results in the reduction of fibrinolysis. Decrease in fibrinolysis process leads to fibrosis and chronicity. Hence, the inhibitory effect of aPC is removed from plasminogen activator inhibitor (PAI), which exerts the inhibitory effect of PAI on plasminogen activator and downstream production of plasmin. Thus, fibrinolysis process mediated by plasmin decreases, and accumulation of fibrin leads to renal damage and chronicity. Decreased level of fibrinogen gamma chain, as a product of fibrinolysis in our dataset (Table 4), supports this hypothesis. The definition of members of coagulation pathway mentioned above has been reviewed by Madhusudhan *et al*[50]. The hypothetical mechanism of renal damage by coagulation and complement pathway is summarized in Figure 5. Furthermore, down-regulation of CD59, an inhibitor of membrane attack complex in complement pathway[53], may result in exacerbation of renal injury.

Decreased excretion of several involved molecules in ECM-receptor interaction pathway in comparison with normal subjects indicates the instability of glycocalyx in the disease process. Glycocalyx injury will affect a broad spectrum of endothelial function including filtration barrier that leads to proteinuria[54]. In addition, a correlation between endothelial cell injury and renal dysfunction in patients with IgAN has been reported[54]. Therefore, it could be suggested that endothelial injury in IgAN may result in the degradation of glycocalyx in kidney tissue, which is recognized by significant changes (down-regulation) of several elements of ECM that contributes to glycocalyx such as CD44, FN1 (fibronectin), SPP1 (OSTP), and HSPG2 (perlecan) (Fig. 4).

In conclusion, functional analysis of the important biomarkers from SLDA and elastic net models revealed that biological processes related to coagulation and acute phase response are involved in the pathogenesis of IgAN. Furthermore, complement and coagulation cascade and ECM-receptor interaction are the two important pathways that impaired in IgAN.

ACKNOWLEDGMENTS

The authors would like to thank the Chronic Kidney Disease Research Center (CKDRC) and Kermanshah University of Medical Sciences for their financial supports. This paper is a part of thesis of Mrs. Mohammadi for MSc. degree.

CONFLICT OF INTEREST. None declared.

REFERENCES

1. Knoppova B, Reily C, Maillard N, Rizk DV, Moldoveanu Z, Mestecky J, Raska M, Renfrow MB, Julian BA, Novak J. The origin and activities of IgA1-containing immune complexes in IgA nephropathy. *Frontiers in immunology* 2016; **7**: 117
2. Sigdel TK, Woo SH, Dai H, Khatri P, Li L, Myers B, Sarwal MM, Lafayette RA. Profiling of autoantibodies in IgA nephropathy, an integrative antibiomics approach. *Clinical Journal of the American Society of Nephrology* 2011; **6**(12): 2775-2784
3. Magistroni R, D D'Agati V, Appel GB, Kiryluk K. New developments in the genetics, pathogenesis, and therapy of IgA nephropathy. *Kidney international* 2015; **88**(5): 974-989.
4. Kalantari S, Nafar M, Rutishauser D, Samavat S, Rezaei-Tavirani M, Yang H, Zubarev RA. Predictive urinary biomarkers for steroid-resistant and steroid-sensitive focal segmental glomerulosclerosis using high resolution mass spectrometry and multivariate statistical analysis. *BMC nephrology* 2014; **15**(1): 141.
5. Kalantari S, Nafar M, Samavat S, Parvin M. ^{1}H

NMR-based metabolomics study for identifying urinary biomarkers and perturbed metabolic pathways associated with severity of IgA nephropathy: a pilot study. *Magnetic Resonance in Chemistry* 2017; **55**(8): 693-699.

6. Kalantari S, Nafar M, Samavat S, Parvin M, Nobakht M.GH BF, Barzi F. ^{1}H NMR-based metabolomics exploring urinary biomarkers correlated with proteinuria in focal segmental glomerulosclerosis: a pilot study. *Magnetic resonance in chemistry* 2016; **54**(10): 821-826.
7. Kalantari S, Nafar M, Samavat S, Rezaei-Tavirani M, Rutishauser D, Zubarev R. Urinary prognostic biomarkers in patients with focal segmental glomerulosclerosis. *Nephro-urology monthly* 2014; **6**(2): e16806.
8. Kalantari S, Rutishauser D, Samavat S, Nafar M, Mahmudieh L, Rezaei-Tavirani M, Zubarev RA. Urinary prognostic biomarkers and classification of IgA nephropathy by high resolution mass spectrometry coupled with liquid chromatography. *PLoS one* 2013; **8**(12): e80830
9. Samavat S, Kalantari S, Nafar M, Rutishauser D, Rezaei-Tavirani M, Parvin M, Zubarev RA. Diagnostic urinary proteome profile for immunoglobulin a nephropathy. *Iranian journal of kidney diseases* 2015; **9**(3): 239-248.
10. Mucha K, Bakun M, Jazwiec R, Dadlez M, Florczak M, Bajor M, Gala K, Paczek L. Complement components, proteolysis-related, and cell communication-related proteins detected in urine proteomics are associated with IgA nephropathy. *Polish archives of internal medicine* 2014; **124**(7-8): 380-386.
11. Neprasova M, Maixnerova D, Novak J, Reily C, Julian BA, Boron J, Novotny P, Suchanek M, Tesar V, Kacer P. Toward noninvasive diagnosis of IgA nephropathy: A pilot urinary metabolomic and proteomic Study. *Disease markers* 2016; **2016**: 3650909.
12. Wang G, Kwan BC, Lai FM, Chow KM, Kam-Tao Li P, Szeto CC. Expression of microRNAs in the urinary sediment of patients with IgA nephropathy. *Disease markers* 2010; **28**(2): 79-86.
13. Berger B, Peng J, Singh M. Computational solutions for omics data. *Nature reviews genetics* 2013; **14**(5): 333-346.
14. Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature reviews cancer* 2008; **8**(1): 37-49.
15. Liu C, Jiang J, Gu J, Yu Z, Wang T, Lu H. High-dimensional omics data analysis using a variable screening protocol with prior knowledge integration (SKI). *BMC systems biology* 2016; **10**(Suppl 4): 118.
16. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the royal statistical society series B (methodological)* 1996; **58**(1): 267-288.
17. Zou H. The adaptive lasso and its oracle properties. *Journal of the American statistical association* 2006; **101**(476): 1418-1429.
18. Ouyang M, Zhang Z, Chen C, Liu X, Liang Y. Application of sparse linear discriminant analysis for metabolomics data. *Analytical methods* 2014; **6**(22): 9037-9044.
19. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the royal statistical sSociety: series B (statistical methodology)* 2005; **67**(2): 301-320.
20. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical association* 2001; **96**(456): 1348-1360.
21. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 1970; **12**(1): 55-67.
22. Cai T, Liu W. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association* 2011; **106**(496): 1566-1577.
23. Merchante LFS, Grandvalet Y, Govaert G. An efficient approach to sparse linear discriminant analysis. Reterieved from https://icml.cc/2012/papers/591.pdf.
24. Qiao Z, Zhou L, Huang JZ. Effective linear discriminant analysis for high dimensional, low sample size data. Reterieved from https://www.stat.tamu.edu/~jianhua /paper/iccsde-sparseLDA.pdf.
25. Shao J, Wang Y, Deng X, Wang S. Sparse linear discriminant analysis by thresholding for high dimensional data. *The annals of statistics* 2011; **39**(2): 1241-1265.
26. Lu J, Plataniotis KN, Venetsanopoulos AN. Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition. *Pattern Recognition Letters* 2005; **26**(2): 181-191.
27. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. *FEMS microbiology reviews* 2011; **35**(2): 343-359.
28. Shahraki HR, Salehi A, Zare N. Survival prognostic factors of male breast cancer in Southern Iran: a LASSO-Cox regression approach. *Asian Pacific journal of cancer prevention* 2015; **16**(15): 6773-6777.
29. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* 2003; **13**(11): 2498-2504.
30. Yokota H, Hiramoto M, Okada H, Kanno Y, Yuri M, Morita S, Naitou M, Ichikawa A, Katoh M, Suzuki H. Absence of increased α_1-microglobulin in IgA nephropathy proteinuria. *Molecular and cellular proteomics* 2007; **6**: 738-744.
31. Yanagawa H, Suzuki H, Suzuki Y, Kiryluk K, Gharavi AG, Matsuoka K, Makita Y, Julian BA, Novak J, Tomino Y. A panel of serum biomarkers differentiates IgA nephropathy from other renal diseases. *PloS one* 2014; **9**(5): e98081
32. Russell MW, Mestecky J, Julian BA, Galla JH. IgA-associated renal diseases: antibodies to environmental antigens in sera and deposition of immunoglobulins and antigens in glomeruli. *Journal of clinical immunology*

1986; **6**(1): 74-86.
33. Kwak NJ, Wang EH, Heo IY, Jin DC, Cha JH, Lee KH, Yang CW, Kang CS, Choi YJ. Proteomic analysis of alpha-1-antitrypsin in immunoglobulin A nephropathy. *Proteomics clinical applications* 2007; **1**(4): 420-428.
34. Moon PG, Lee JE, You S, Kim TK, Cho JH, Kim IS, Kwon TH, Kim CD, Park SH, Hwang D, Kim YL, Baek MC. Proteomic analysis of urinary exosomes from patients of early IgA nephropathy and thin basement membrane nephropathy. *Proteomics* 2011; **11**(12): 2459-2475.
35. Sedic M, Gethings LA, Vissers JP, Shockcor JP, McDonald S, Vasieva O, Lemac M, Langridge JI, Batinić D, Pavelić SK. Label-free mass spectrometric profiling of urinary proteins and metabolites from paediatric idiopathic nephrotic syndrome. *Biochemical and biophysical research communications* 2014; **452**(1): 21-26.
36. Khan TN, Sinniah R. Renal tubular antiproteinase (alpha-1-antitrypsin and alpha-1-antichymotrypsin) response in tubulo-interstitial damage. *Nephron* 1993; **65**(2): 232-239.
37. Candiano G, Musante L, Bruschi M, Petretto A, Santucci L, Del Boccio P, Pavone B, Perfumo F, Urbani A, Scolari F, Ghiggeri GM. Repetitive fragmentation products of albumin and α1-antitrypsin in glomerular diseases associated with nephrotic syndrome. *Journal of the American society of nephrology* 2006; **17**(11): 3139-3148.
38. Graziadei I, Weiss G, Bohm A, Werner-Felmayer G, Vogel W. Unidirectional upregulation of the synthesis of the major iron proteins, transferrin-receptor and ferritin, in HepG2 cells by the acute-phase protein α1-antitrypsin. *Journal of hepatology* 1997; **27**(4): 716-725.
39. Pike RN, Buckle AM, Le Bonniec BF, Church FC. Control of the coagulation system by serpins. *FEBS journal* 2005; **272**(19): 4842-4851.
40. Machii R, Sakatume M, Kubota R, Kobayashi S, Gejyo F, Shiba K. Examination of the molecular diversity of α1 antitrypsin in urine: deficit of an α1 globulin fraction on cellulose acetate membrane electrophoresis. *Journal of clinical laboratory analysis* 2005; **19**(1): 16-21.
41. Yanagisawa H, Davis EC, Starcher BC, Ouchi T, Yanagisawa M, Richardson JA, Olson EN. Fibulin-5 is an elastin-binding protein essential for elastic fibre development *in vivo*. *Nature* 2002; **415**(6868): 168-171.
42. Prikryl P, Vojtova L, Maixnerova D, Vokurka M, Neprasova M, Zima T, Tesar V. Proteomic approach for identification of IgA nephropathy-related biomarkers in urine. *Physiological research* 2017; **66**(4): 621-632.
43. Ohara H, Akatsuka S, Nagai H, Liu YT, Jiang L, Okazaki Y, Yamashita Y, Nakamura T, Toyokuni S. Stage-specific roles of fibulin-5 during oxidative stress-induced renal carcinogenesis in rats. *Free radical research* 2011; **45**(2): 211-220.
44. Sullivan KM, Bissonnette R, Yanagisawa H, Hussain SN, Davis EC. Fibulin-5 functions as an endogenous angiogenesis inhibitor. *Laboratory investigation* 2007; **87**(8): 818-827.
45. Matafora V, Cuccurullo M, Beneduci A, Petrazzuolo O, Simeone A, Anastasio P, Mignani R, Feriozzi S, Pisani A, Comotti C, Bachi A, Capasso G. Early markers of Fabry disease revealed by proteomics. *Molecular biosystems* 2015; **11**(6): 1543-1551.
46. Gang X, Ueki K, Kon S, Maeda M, Naruse T, Nojima Y. Reduced urinary excretion of intact osteopontin in patients with IgA nephropathy. *American journal of kidney diseases* 2001; **37**(2): 374-379.
47. Bautista DS, Denstedt J, Chambers AF, Harris JF. Low-molecular-weight variants of osteopontin generated by serine proteinases in urine of patients with kidney stones. *Journal of cellular biochemistry* 1996; **61**(3): 402-409.
48. Sano N, Kitazawa K, Sugisaki T. Localization and roles of CD44, hyaluronic acid and osteopontin in IgA nephropathy. *Nephron* 2001; **89**(4): 416-421.
49. Janssen U, Bahlmann F, Köhl J, Zwirner J, Haubitz M, Floege J. Activation of the acute phase response and complement C3 in patients with IgA nephropathy. *American journal of kidney diseases* 2000; **35**(1): 21-28.
50. Madhusudhan T, Kerlin BA, Isermann B. The emerging role of coagulation proteases in kidney disease. *Nature reviews nephrology* 2016; **12**(2): 94-109.
51. Matsubara M, Akiu N, Ootaka T, Saito T, Yoshinaga K. Glomerular deposition of coagulation factors VII, VIII, and IX in IgA nephropathy: possible coagulation system involvement in IgA nephropathy. *Nephron* 1989; **53**(4): 381-383.
52. Rabelink TJ, De Zeeuw D. The glycocalyx–linking albuminuria with renal and cardiovascular disease. *Nature reviews nephrology* 2015; **11**(11): 667-676.
53. Alegretti AP, Mucenic T, Brenol JCT, Xavier RM. The role of CD55/CD59 complement regulatory proteins on peripheral blood cells of systemic lupus erythematosus patients. *Revista brasileira de reumatologia* 2009; **49**(3): 276-287
54. Kusano T, Takano H, Kang D, Nagahama K, Aoki M, Morita M, Kaneko T, Tsuruoka S, Shimizu A. Endothelial cell injury in acute and chronic glomerular lesions in patients with IgA nephropathy. *Human pathology* 2016; **49**: 135-144.