Nonlinear Sparse Component Analysis with a Reference: Variable Selection in Genomics and Proteomics

Ivica Kopriva[1(✉)], Sanja Kapitanović[2], and Tamara Čačev[2]

[1] Division of Laser and Atomic R&D, Ruđer Bošković Institute, Bijenička Cesta 54, 10000 Zagreb, Croatia
Ivica.Kopriva@irb.hr
[2] Division of Molecular Medicine, Ruđer Bošković Institute, Bijenička Cesta 54, 10000 Zagreb, Croatia
{Sanja.Kapitanovic,Tamara.Cacev}@irb.hr

Abstract. Many scenarios occurring in genomics and proteomics involve small number of labeled data and large number of variables. To create prediction models robust to overfitting variable selection is necessary. We propose variable selection method using nonlinear sparse component analysis with a reference representing either negative (healthy) or positive (cancer) class. Thereby, component comprised of cancer related variables is automatically inferred from the geometry of nonlinear mixture model with a reference. Proposed method is compared with 3 supervised and 2 unsupervised variable selection methods on two-class problems using 2 genomic and 2 proteomic datasets. Obtained results, which include analysis of biological relevance of selected genes, are comparable with those achieved by supervised methods. Thus, proposed method can possibly perform better on unseen data of the same cancer type.

Keywords: Variable selection · Nonlinear mixture model · Empirical kernel maps · Sparse component analysis

1 Introduction

Data acquired by microarray gene expression profiling technology [1, 2] or mass spectrometry [3, 4], present "large p, small *n*" problem: large number of variables (genes or mass-to-charge, *m/z*, ratios) and small number of labeled (diagnosed) gene or protein expressions. They correspond with the mixtures in blind source separation (BSS) vocabulary while variables correspond with samples in BSS vocabulary. In described scenario learned prediction models adapt to training data (overfitt) and not generalize well on unseen data of the same cancer type [5, 6]. Improvement of predictor performance is enabled by variable selection [6, 7]. This implies selection of small number of variables that discriminate well between cancer and healthy subjects. Here we propose unsupervised variable selection method that performs blind sparseness constrained decomposition of each mixture independently according to implicit, empirical kernel map (EKM)-based [8], nonlinear mixture model. The model is comprised of a test mixture and a reference mixture representing positive (cancer) class.

E. Vincent et al. (Eds.): LVA/ICA 2015, LNCS 9237, pp. 168–175, 2015.
DOI: 10.1007/978-3-319-22482-4_19

Proposed method takes into account biological diversity of mixtures as well as nonlinear nature of the interaction among variables (genes) within the components present in mixtures [9]. Reference mixture enables automatic selection of component within test mixture that is comprised of cancer related variables. Since no label information is used selected cancer related components can be used both for biomarker identification studies as well as for training prediction models. As opposed to that, variable selection based on standard BSS methods, [10–14], use whole dataset for decomposition. Afterwards, one component composed of cancer related variables is selected by using label information. That enables selected component to be used for biomarker identification studies but prevents it to be used for training predictive models (otherwise label information would be used twice). The method proposed herein is nonlinear generalization of the mixture dependent linear model with a reference [15] as well as generalization of mixture dependent nonlinear model with a reference that is based on approximate explicit feature maps (EFM) [16]. Implicit nonlinear mapping is performed variable-wise yielding nonlinear mixture model with the same number of variables and "increased" number of mixtures. Sparse component analysis (SCA) is performed on nonlinearly mapped mixture. Afterwards, variables in cancer related components are ranked by their mixture-wise variance. That yields index set used to access variables in the original input space. They are used to learn two-class support vector machine (SVM) predictive model [17]. We compare proposed method with 3 supervised variable selection methods [18, 19] and 2 unsupervised methods [15, 16]. The methods were compared on 2 well-known cancer types in genomics: colon cancer [1] and prostate cancer [2], as well as on 2 well-known cancer types in proteomics: ovarian cancer [3] and prostate cancer [4]. Furthermore, analysis of biological relevance of selected genes in colon cancer experiment is also provided. We describe proposed method in Sect. 2. Results of comparative performance analysis are described in Sect. 3. Conclusions are proposed in Sect. 4.

2 Method

Let us assume that N mixtures are stored in rows of data matrix $\mathbf{X} \in \mathbb{R}^{N \times K}$, whereas each mixtures is further comprised of K variables. We also assume that N mixtures have diagnoses (label): $\mathbf{x}_n \in \mathbb{R}^{1 \times K}, y_n \in \{1, -1\}$, $n = 1,\ldots, N$, where 1 stands for positive (cancer) and -1 stands for negative (healthy) mixture. Within this paper we assume that mixtures are normalized such that: $-1 \leq x_{nk} \leq 1 \, \forall n = 1, \ldots, N$ $k = 1, \ldots, K$. Matrix factorization methods such as principal component analysis, independent component analysis, SCA and/or nonnegative matrix factorization assume linear mixture model: $\mathbf{X} = \mathbf{AS}$, where $\mathbf{A} \in \mathbb{R}_{0+}^{N \times M}$, $\mathbf{S} \in \mathbb{R}^{M \times K}$ and M stands for an unknown number of components imprinted in mixtures. Each component is represented by a row vector of matrix $\mathbf{S}$, that is: $\mathbf{s}_m \in \mathbb{R}^{1 \times K}$, $m = 1,\ldots, M$. Column vectors of matrix $\mathbf{A}$: $\mathbf{a}_m \in \mathbb{R}^{N \times 1}$, $m = 1,\ldots, M$, represent concentration profiles of the corresponding components. To infer component comprised of disease relevant variables label information is used by methods such as [10, 11]. That prevents usage of selected component for training prediction models. This limitation has been addressed in [15] by

formulating mixture dependent linear model with a reference. Herein, as in [16], we assume nonlinear model of a mixture:

$$\begin{bmatrix} x_{ref,k} \\ x_{nk} \end{bmatrix} = f_n(\mathbf{s}_{k;n}) \;\; n = 1,\ldots,N;\; k = 1,\ldots,K \tag{1}$$

where $f_n : \mathbb{R}^{M_n} \rightarrow \mathbb{R}^2$ is an unknown mixture dependent nonlinear function that maps M_n-dimensional vector of variables $\mathbf{s}_{k;n} \in \mathbb{R}^{M_n \times 1}$ to 2-dimensional observation vector. Thereby, first element of the observation vector belongs to the reference mixture and second element to the test mixture. Herein, we assume that reference mixture represents positive (cancer) class. It can be selected by an expert or, as it was the case herein, can be obtained by averaging all the mixtures belonging to positive class. We propose EKM for implicit (kernel-based) mapping of (1). We repeat definition 2.15 from [8]:

Definition 1. For a given set of patterns $\{\mathbf{v}_d \in \mathbb{R}^{N\times 1}\}_{d=1}^{D} \subset \mathbf{X}$, $D \in \mathbb{N}$, we call $\psi : \mathbb{R}^N \rightarrow \mathbb{R}^D$, where $\psi :\; \mathbf{x}_{nk} \mapsto [\kappa(\mathbf{v}_1, \mathbf{x}_{nk}),\ldots,\kappa(\mathbf{v}_D, \mathbf{x}_{nk})]^T$, $\forall k = 1,\ldots,K$, the EKM with respect to basis $\mathbf{V} := \{\mathbf{v}_d\}_{d=1}^{D}$.

Thereby, $\mathbf{x}_{nk} = \begin{bmatrix} x_{ref,k} & x_{nk} \end{bmatrix}^T$ is defined in (1). The basis $\mathbf{V}$ has to satisfy:

$$span\{\mathbf{v}_d\}_{d=1}^{D} \approx span\{\mathbf{x}_{nk}\}_{k=1}^{K} \tag{2}$$

To estimate $\mathbf{V}$ we have used *k-means* algorithm to cluster empirical set of patterns (samples) $\{\mathbf{x}_{nk}\}_{k=1}^{K}$ in predefined number of D cluster centroids (basis vectors). If $\mathbf{V}$ satisfies (2) then obviously $\mathbf{V} \cup [1 \;\; 0]^T$ satisfies (2) as well. Hence, EKM $\psi(\mathbf{x}_{nk})$ is obtained by projecting EFM $\phi(\mathbf{x}_{nk})$ associated with kernel $\kappa(\circ, \mathbf{x}_{nk})$ on a $(D + 1)$-dimensional subspace in mapping induced space spanned by $\left\{\phi(\mathbf{v}_d) \in \mathbb{R}^{\bar{D}}\right\}_{d=1}^{D+1}$:

$$\begin{aligned} \psi(\mathbf{x}_{nk}) &= [\phi(\mathbf{v}_1) \;\ldots\; \phi(\mathbf{v}_D)\; \phi(\mathbf{v}_{D+1}) \;]^T \phi(\mathbf{x}_{nk}) \\ &= [\kappa(\mathbf{x}_{nk}, \mathbf{v}_1) \;\ldots\; \kappa(\mathbf{x}_{nk}, \mathbf{v}_D) \;\; \kappa(\mathbf{x}_{nk}, \mathbf{v}_{D+1})]^T \quad \forall k = 1,\ldots,K \end{aligned} \tag{3}$$

where $\mathbf{v}_{D+1} \in \mathbb{R}_{0+}^{2\times 1} = [1\; 0]^T$. We now define mixture dependent linear model in EKM-induced space:

$$\psi\begin{pmatrix} x_{ref,k} \\ x_{nk} \end{pmatrix} \approx \bar{\mathbf{A}}_n \bar{\mathbf{s}}_{k;n} \quad k = 1,\ldots,K \tag{4}$$

where $\bar{\mathbf{A}}_n \in \mathbb{R}_{0+}^{D+1\times M_n}$, $\bar{\mathbf{s}}_{k;n} \in \mathbb{R}^{M_n \times 1}$ and M_n stands for mixture dependent number of components. The key observation regarding nonlinear model (3)/(4) is that, for suitably chosen kernel, $\kappa(\mathbf{x}_{nk}, \mathbf{v}_{D+1})$ it becomes a function of the reference mixture $x_{ref,k}$ only. As an example, for $\kappa(\mathbf{x}_{nk}, \mathbf{v}_{D+1}) = \exp\left(|\langle \mathbf{x}_{nk}, \mathbf{v}_{D+1}\rangle|/\sigma^2\right) = \exp\left(x_{ref,k}/\sigma^2\right)$. For Gaussian kernel it applies: $\kappa(\mathbf{x}_{nk}, \mathbf{v}_{D+1}) = \exp(-x_{nk}^2/\sigma^2)\exp(-\sigma^2)\exp\left(\left(2x_{ref,k} - x_{ref,k}^2\right)/\sigma^2\right)$.

Under assumption $-1 \le x_{nk} \le 1$ and $\sigma^2 > x_{nk}^2$ the first part is approximately 1 and the last part $\exp\left((2x_{ref,k} - 1)/\sigma^2\right)$. Thus, $\kappa(\mathbf{x}_{nk}, \mathbf{v}_{D+1}) \approx \exp\left((2x_{ref,k} - 1)/\sigma^2\right)$. Hence, we can express $\psi(\mathbf{x}_{nk})$ in standard Euclidean basis $\{\mathbf{e}_d\}_{d=1}^{D+1}$:

$$\psi\begin{pmatrix} x_{ref,k} \\ x_{nk} \end{pmatrix} = \kappa(\mathbf{x}_{nk}, \mathbf{v}_1)\mathbf{e}_1 + \ldots + \kappa(\mathbf{x}_{nk}, \mathbf{v}_D)\mathbf{e}_D + f\left(x_{ref,k}\right)\mathbf{e}_{D+1} \tag{5}$$

Representation (5) enables automatic selection of component $\bar{\mathbf{s}}_{m*}$, $m^* \in \{1,.., M_n\}$ comprised of cancer relevant variables. $\bar{\mathbf{s}}_{m*}$ is associated with the mixing vector that closes the smallest angle with the axis $\mathbf{e}_{D+1}$ that represents cancer class. Cosine of the angle between mixing vector $\bar{\mathbf{a}}_{m;n}$ and $\mathbf{e}_{D+1}$ s obtained as:

$$\cos\angle\left(\bar{\mathbf{a}}_{m;n}, \mathbf{e}_{D+1}\right) = \left\langle \bar{\mathbf{a}}_{m;n}, \mathbf{e}_{D+1} \right\rangle / \left\| \bar{\mathbf{a}}_{m;n} \right\| \tag{6}$$

Thus index of component composed of cancer relevant variables is obtained as:

$$m^* = \arg\max_m \cos\angle\left(\bar{\mathbf{a}}_{m;n}, \mathbf{e}_{D+1}\right) \tag{7}$$

When each mixture is decomposed according to (4), components comprised of cancer relevant variables are stored row-wise in a matrix $\bar{\mathbf{S}}_{cancer} \in \mathbb{R}^{N \times K}$. Variables (columns of $\bar{\mathbf{S}}_{cancer}$) are then ranked by their variance across the mixture dimension yielding $\bar{\mathbf{S}}_{cancer}^{ranked} \in \mathbb{R}^{N \times K}$. Let us denote by I a corresponding index set. Variables ranked in the original mixture space are obtained by indexing each mixture by I, that is: $\mathbf{x}_n^{ranked} = \mathbf{x}_n(I)$, $n = 1,\ldots, N$. Mixtures with ranked variables form rows of the matrix $\mathbf{X}^{ranked} \in \mathbb{R}^{N \times K}$. That, when paired with the vector of labels $\mathbf{y}$, is used to learn SVM prediction model.

Decomposition of the linear mixture model (4) is performed enforcing sparseness of the components $\bar{\mathbf{s}}_{m;n}$, $m = 1, \ldots, M_n$. That is because sparse components are comprised of few dominantly expressed variables and that can be good indicator of a disease. Method used to solve, in principle, underdetermined BSS problem (4) estimates mixing matrix $\bar{\mathbf{A}}_n$ first by using the separable NMF algorithm [20] with a MATAB code available at: https://sites.google.com/site/nicolasgillis/publications. The important characteristic of the method [20] is that there are no free parameters to be tuned or defined a priori. The unknown number of components M_n is also estimated automatically and is limited above by $D + 1$. Thus, by cross-validating D we implicitly cross-validate M_n as well. After $\bar{\mathbf{A}}_n$ is estimated, $\bar{\mathbf{S}}_n$ is estimated by solving sparseness constrained optimization problem:

$$\hat{\bar{\mathbf{S}}}_n = \min_{\bar{\mathbf{S}}_n} \left\{ \frac{1}{2} \left\| \hat{\bar{\mathbf{A}}}_n \bar{\mathbf{S}}_n - \psi\begin{pmatrix} \mathbf{x}_{ref} \\ \mathbf{x}_n \end{pmatrix} \right\|_F^2 + \lambda \left\| \bar{\mathbf{S}}_n \right\|_1 \right\} \tag{8}$$

where the hat sign denotes an estimate of the true (but unknown) quantity, λ is regularization parameter and $\left\| \bar{\mathbf{S}}_n \right\|_1$ denotes ℓ_1-norm of $\bar{\mathbf{S}}_n$. To solve (8), we have used the

iterative shrinkage thresholding (IST) method [21] with a MATLAB code at: http://ie.technion.ac.il/Home/Users/becka.html. Sparsity of the solution is controlled by the parameter λ. There is a maximal value of λ (denoted by λ_{max} here) above which the solution of the problem (8) is equal to zero. Thus, in the experiments reported below the value of λ has been selected by cross-validation with respect to λ_{max}.

3 Experiments

Proposed approach is compared with supervised variable selection methods: maximum mutual information minimal redundancy (MIMR) method [18] and HITTON_PC and HITTON_MB [19] methods. We also report results for linear [15] and EFM-based nonlinear [16] counterparts of proposed method. Gene Expression Model Selector (GEMS) software system [22], has been used for 10-fold cross-validation based learning of SVM-based diagnostic models with polynomial and Gaussian kernels. The system is available at: http://www.gems-system.org/. HITON_PC and HITON_MB algorithms are implemented in GEMS software system while implementation of the MIMR algorithm is available at MATLAB File Exchange. Order D of the EKM in (3) has been cross-validated in the range: $D \in \{5, 10, 15, 20, 25, 30\}$. Regularization constant λ in (8) has been cross-validated in the range: $\lambda \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\} \times \lambda_{max.}$ Methods were compared on 2 cancer types in genomics: colon cancer [1] and prostate cancer [2], as well as on 2 cancer types in proteomics: ovarian cancer [3] and prostate cancer [4]. The number of cancer vs. normal mixtures is for 4 datasets given in respective order as: 40/22, 52/50, 100/100 and 69/63. The number of variables in each dataset is in respective order given as: 2000, 10509, 15152 and 15154. For each dataset we report in Table 1 result achieved by: proposed method, the best result achieved by one of 3 supervised methods and results achieved by [15, 16]. Due to the

Table 1. Classification accuracy and number of selected variables.

Dataset	Proposed method	Supervised method	[16]	[15]
1. Prostate cancer	91.18 % / 12 genes (D = 20, λ = 0.3, Gauss kernel).	MIMR: **98.08 % / 10 genes**	91.27 % / 38 genes	94.27 % / 477 genes.
2. Colon cancer	**93.57 %** / 20 genes (D = 20, λ = 0.3, Gauss kernel).	HITON_MB: 93.33 % / **4 genes**	91.91 % / 24 genes	90.48 % / 30 genes, λ = 0.05.
3. Ovarian cancer	94.5 % / 47 *m/z* lines (D = 20, λ = 0.35, Exp. kernel).	HITON_PC: **99.5 % / 7 *m/z* lines**	93 % / 7 *m/z* lines	82 % / 25 *m/z* lines, λ = 0.2.
4. Prostate cancer	94.61 % / 27 *m/z* lines (D = 20, λ = 0.35, Exp. kernel).	MIMR: **100 % / 9 *m/z* lines**	94.06 % / 14 *m/z* lines	94.01 % / 85 *m/z* lines, λ = 0.2.

lack of space we do not report details on parameters of the SVM classifiers. For each of 4 datasets, proposed method achieves result that is worse than but comparable with the result of supervised algorithm and better than its linear and EFM-based nonlinear unsupervised counterparts [15, 16]. Since reported results are achieved with small number of variables the probability of overfitting is reduced. Thus, it is reasonable to expect that performance on unseen data of the same cancer type by proposed unsupervised method will be better than the one achieved with supervised algorithms.

Colon cancer data are available at: http://genomic-pubs.princeton.edu/oncology/affydata/index.html. Prostate cancer genomic data are available at: http://www.gems-system.org/. Ovarian and prostate cancer proteomic data are available at: http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp. To comply with principle of reproducible research software which implements proposed algorithm, datasets used and results presented in Table 1 are available at: http://www.lair.irb.hr/ikopriva/Data/HRZZ/data/LVA_2015.zip.

We also provide brief biological interpretation of genes selected by proposed method in the colon cancer experiment [1]. The majority of genes selected by the proposed algorithm have been previously associated with tumorgenesis. For instance, expression of genes encoding ribosomal proteins (RPS9, RPS18, RPS29, RPS24, RPLP1, RPL30) has been known to increase in tumors as a result of uncontrolled cell proliferation which is one of the key hallmarks of cancer. In addition, several previous microarray studies have reported an increase in mRNA expression of ribosomal genes in solid tumors including colorectal cancer [23]. Several genes which were found to be differentially expressed by our algorithm like IGHG3, FTL, GAPDH and UBC encode proteins involved in cellular metabolism and bioenergetics and have previously been associated with cancer [24, 25]. This is not surprising since changes in metabolic processes are often observed in tumor cells. For instance altered GAPDH expression has been reported in breast, gastric, liver, lung as well as colorectal cancer [26]. Laminin receptor 1 (RPSA) and actin (ACTB), two other genes detected by our algorithm, are involved in wide spectrum of cellular functions including the maintenance of cellular structure as well as adhesion and motility [26]. When specifically colorectal cancer is considered, S100A6 has previously been associated with this type of cancer [27]. In addition, the role of Thymosin beta-4 in cell proliferation, growth and migration has been previously established and its overexpression has been reported during the different stages of colorectal carcinogenesis [28].

4 Conclusion

Since it requires little prior knowledge unsupervised decomposition of a set of mixtures into additive combination of components is of particular importance in addressing overfitting problem. Herein, we have proposed an unsupervised approach for variable selection by decomposing each mixture individually into sparse components according to nonlinear kernel-based model of a mixture, whereas decomposition is performed with respect to a reference mixture that represents positive (cancer) class. That enables selection of cancer related components automatically and, afterwards, their use for either biomarker identification studies or learning diagnostic models. It is conjectured

that outlined properties of proposed method enabled competitive diagnostic accuracy with small number of variables on cancer related human gene and protein expression datasets. While proposed method is developed for binary (two-class) problems its extension for multi-category classification problems is aimed for the future work.

Acknowledgments. Work of I. Kopriva has been partially supported through the FP7-REGPOT-2012-2013-1, Grant Agreement Number 316289 – InnoMol and partially through the Grant 9.01/232 funded by Croatian Science Foundation.

References

1. Alon, U., et al.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. U.S.A. **96**, 6745–6750 (1999)
2. Singh, D., et al.: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell **1**, 203–209 (2002)
3. Petricoin, E.F., et al.: Use of proteomic patterns in serum to identify ovarian cancer. Lancet **359**, 572–577 (2002)
4. Petricoin, E.F., et al.: Serum proteomic patterns for detection of prostate cancer. J. Natl. Canc. Inst. **94**, 1576–1578 (2002)
5. Guyon, I., et al.: Gene selection for cancer classification using support vector machines. Mach. Learn. **46**, 389–422 (2002)
6. Statnikov, A., et al.: A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics **21**, 631–643 (2005)
7. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2002)
8. Schölkopf, B., Smola, A.: Learning with Kernels. The MIT Press, Cambridge (2002)
9. Yuh, C.H., Bolouri, H., Davidson, E.H.: Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. Science **279**, 1896–1902 (1998)
10. Lee, S.I., Batzoglou, S.: Application of independent component analysis to microarrays. Genome Biol. **4**, R76 (2003)
11. Schachtner, R., et al.: Knowledge-based gene expression classification via matrix factorization. Bioinformatics **24**, 1688–1697 (2008)
12. Stadtlthanner, K., et al.: Hybridizing sparse component analysis with genetic algorithms for microarray analysis. Neurocomputing **71**, 2356–2376 (2008)
13. Gao, Y., Church, G.: Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics **21**, 3970–3975 (2005)
14. Kim, H., Park, H.: Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics **23**, 1495–1502 (2007)
15. Kopriva, I., Filipović, M.: A mixture model with a reference-based automatic selection of components for disease classification from protein and/or gene expression levels. BMC Bioinformatics **12**, 496 (2011)
16. Kopriva, I.: A Nonlinear Mixture Model Based Unsupervised Variable Selection in Genomics and Proteomics. In: Bioinformatics 2015 – 6th International Conference on Bioinformatics Models, Methods and Algorithms, pp. 85–92, Scitepress (2015)
17. Vapnik, V.: Statistical Learning Theory. Wiley-Interscience, New York (1998)

18. Brown, G.: A new perspective for information theoretic feature selection. J. Mach. Learn. Res. **5**, 49–56 (2009)
19. Aliferis, C.F., et al.: Local causal and markov blanket induction for causal discovery and feature selection for classification - Part I: algorithms and empirical evaluation. J. Mach. Learn. Res. **11**, 171–234 (2010)
20. Gillis, N., Vavasis, S.A.: Fast and robust recursive algorithms for separable nonnegative matrix factorization. IEEE Trans. Pattern Anal. Mach. Intell. **36**, 698–714 (2014)
21. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. **2**, 183–202 (2009)
22. Statnikov, A., et al.: GEMS: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. Int. J. Med. Inf. **74**, 491–503 (2003)
23. Artero-Castro, A., et al.: Rplp1 bypasses replicative senescence and contributes to transformation. Exp. Cell Res. **315**, 1372–1383 (2009)
24. Bin Amer, S.M., et al.: Gene expression profiling in women with breast cancer in a Saudi population. Saudi Med. J. **29**, 507–513 (2008)
25. Alkhateeb, A.A., Connor, J.R.: The significance of ferritin in cancer: anti-oxidation, inflammation and tumorigenesis. Biochim. Biophys. Acta **1836**, 245–254 (2013)
26. Guo, C., Liu, S., Sun, M.Z.: Novel insight into the role of GAPDH playing in tumor. Clin. Transl. Oncol. **15**, 167–172 (2013)
27. Leśniak, W., Słomnicki, Ł.P., Filipek, A.: S100A6 - new facts and features. Biochem Biophys Res Commun. **390**, 1087–1092 (2009)
28. Sribenja, S., et al.: Roles and mechanisms of β-thymosins in cell migration and cancer metastasis: an update. Cancer Invest. **31**, 103–110 (2013)