Reconstructing Gene Regulatory Network Using Heterogeneous Biological Data

Farzana Kabir Ahmad and Nooraini Yusoff

Bio-Inspired Agent Systems, Artificial Inelligence Lab, School of Computing,
College of Arts and Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah Malaysia
{farzana58,nooraini}@uum.edu.my

Abstract. Gene regulatory network is a model of a network that describes the relationships among genes in a given condition. However, constructing gene regulatory network is a complicated task as high-throughput technologies generate large-scale of data compared to number of sample. In addition, the data involves a substantial amount of noise and false positive results that hinder the downstream analysis performance. To address these problems Bayesian network model has attracted the most attention. However, the key challenge in using Bayesian network to model GRN is related to its learning structure. Bayesian network structure learning is NP-hard and computationally complex. Therefore, this research aims to address the issue related to Bayesian network structure learning by proposing a low-order conditional independence method. In addition we revised the gene regulatory relationships by integrating biological heterogeneous dataset to extract transcription factors for regulator and target genes. The empirical results indicate that proposed method works better with biological knowledge processing with a precision of 83.3% in comparison to a network that rely on microarray only, which achieved correctness of 80.85%.

Keywords: Gene regulatory, Bayesian network, heterogeneous data.

1 Introduction

DNA microarray is one of the most fascinating and latest breakthrough technologies in molecular biology. This technology has been used to facilitate the quantitative studies of thousand of genes in order to answer various research questions. To date, this technology is widely employed to construct gene regulatory network (GRN). The construction of GRN using microarray data has enabled the measurement of global response of biological system to examine specific inventions. For example, scientists can look into large number of gene interactions that are perturbed during cancer progression.

GRN also known as cellular network is a set of molecular components that includes genes, proteins and other molecules, which collectively accomplish cellular functions as these molecules interact with each other [1]. However, in this study we only used microarray data to model gene network. The fundamental idea behind GRN

S. Ramanna et al. (Eds.): MIWAI 2013, LNCS 8271, pp. 97–107, 2013.

analysis is to discover regulator genes by examining gene expression patterns. Notably, some genes regulate other genes, which mean that the amount of a gene expressed at a certain time could activate or inhibit the expression of another gene. Thus, the regulation of gene expression has an important role in cellular functions. Changes in the expression levels of particular genes across a whole process, such as the cell cycle or response to certain treatments, have provided information that allows reconstruction of cellular network using reverse engineering technique.

A large number of works have reported that GRN can possibly assist researchers in suggesting and evaluating innovative hypotheses in the context of genetic regulatory processes [2-3]. Such data-driven regulatory networks analysis ultimately would provide clearer understanding of the genetic regulatory processes, which are normally complex and intricate. Furthermore, it would bring significant implications in the biomedical fields and many other pharmaceutical industries. Thus, identifying GRNs and understanding regulatory processes at the genetic level has become an imperative goal in computational biology.

Various mathematical and computational methods have been used to model GRN from microarray data, including Boolean network, pair-wise comparison, differential equations estimation, Bayesian network [4-5] and other techniques. Amongst these, the Bayesian network model attracts the most attention and has become the prominent technique because it can capture linear, nonlinear, combinatorial, stochastic and casual relationships between variables. Compared to other methods, Bayesian network model establishes considerable relationships between all genes in the system. In addition, due to rich probabilistic semantics, this model is also capable of working with noisy data that is a common problem in microarray data. Furthermore, this technique allows for different implicit variable information to be added to the networks, which possibly enhances the interpretation of the gene regulation process. Thus, Bayesian network is used in this study to analyze gene regulatory processes and to model gene relationships for breast cancer metastasis.

The key challenge in using Bayesian network to model GRN is related to it's structure learning. Bayesian network structure learning is NP-hard and computationally complex, as the number of possible graphs increases super-exponentially with the number of genes and an exhaustive search is untraceable. This difficulty is a common problem in gene regulatory analysis because network is usually learnt from a relatively small number of measurements. The high dimensionalities of microarray data, which usually contain insufficient sample measurement plus large number of genes to examine, are the main causes of this problem.

The basic idea is to develop GRN by measuring the dependencies among nodes of the given data. Low-order conditional independence is used to examine the relationships between genes. Although the proposed method has increased the accuracy of inferred network as reported in [6], such gene network is solely based on the microarray data and is often insufficient for rigorous analysis. In many cases, microarray data is often daunted by noisy, incomplete data and misleading outliers, which can produce high number of false positive edges. Accordingly, an inferred GRN may contain some incorrect gene regulations that are unreliable from the biological point of view. Thus, integration of biological knowledge into gene network has become necessary to overcome the problem.

This study has used heterogeneous biological data to improve the structure learning of Bayesian network. The remainder of this paper is organized as follows. Section 2 describes some previous works, whose have utilized biological data to achieve better construction of GRN. Section 3 on the other hand, presents the proposed method. Section 4 meanwhile presents experimental results and discussion. Finally, Section 5 offers concluding and future direction remarks.

2 Previous Works

Recent years have witnessed the increasing amount of genomic data such as gene expression, single-nucleotide polymorphism (SNP) and proteomic which are available in public databases. This trend has triggered a new research direction whereby researchers now are motivated to combine various kinds of genomic data to reconstruct GRN. However, the integration of different data source is not simple as these data varies in term of sizes, formats and types. Furthermore, most of these data is partly independent and provide complementary information on the whole genome. Since, there is no complete GRN that are available for any species, the best option in hand is to integrate diverse biological data that presents fragmented information and seek a better explanation for the development at a system level.

Data integration has been defined as a data fusion process that not only includes various data sources but also provides biological meaning with the use of bioinformatics and computational tools. The overreaching goal of data integration is to obtain more accurate, precise and broader view of network than any of single dataset. Based on this concept several works have been done to seek for better explanation of GRN (as explained in Section 2.1. and Section 2.2). Generally there are two types of data integration in the field of GRN, namely homologous data integration and heterogeneous data integration. Homologous data integration mainly uses of similar data type for example combination of multiple microarray datasets from different studies to answer question raised by researchers. Meanwhile, heterogeneous data integration make used of different data types across or within studies to seek for better clarify of information provided by a single data type.

The main idea for homologous data integration is to increase the number of samples to address the issue of high-dimensional data. Most studies in homologous data integration have focused on comparing two or more related datasets to identify significant genes that can distinguish different group of samples (e.g. disease and normal samples). For an example, Rhodes *et al.* [7] have combined multiple microarray datasets to classify common transcription profiles that are universally activated in most cancer types.

Unlike homologous data integration, where it used similar data types, heterogeneous data integration mainly focuses on applying various data sources to ensure the reliability of results obtained. Among the popular data integration is gene expression and proteomic data. Protein is the end product of translation process and is also used as a trigger to initiate the expression of other genes. Therefore, the combination of these data type is reasonable to most of researchers. Thus, many previous works have estimated co-expressed relationship as a gene regulatory instead of looking at protein-protein interactions [8-9]. Besides that, large number of researchers also utilized transcription factor binding sites (TFBS) to verify the GRN. Like protein, TFBS is another complementary data to measure cellular state. Hence, more recent works have explored data integration of external knowledge to identify transcription factors and their target genes [10-11]. Transcription factors are very essential in regulating gene expression. Motivated by this fundamental concept, transcription factors have been used in this research to discover significant biological information from high-throughput data.

3 Methods

The Bayesian network is a graphical model that was introduced by Pearl and Wright in 1980s [12]. To deal with a large number of genes in microarray data, this research defines the Bayesian network, BN as: $BN = (G, P)$ where $G = (X, E(G))$ is a DAG with a set of variables X representing $\{X_i;\ i \in V\}$, and $E(G) \subseteq X_i * X_j$ (set of pairs that represents the dependent among v variables). The element E is an edge from node X_i to X_j, indicating X_i is a parent to X_j. On the other hand, P corresponds to joint distribution on the variables in the network. The $Pa(V)$ represents the parent for a set of vertex V and can be defined as:

$$Pa(X_i, G) = \{X_j,\ such\ that\ (X_i, X_j) \in E(G);\ j \in V\} \qquad (1)$$

where $Pa(X_i, G)$ is the parent of X_i in the graph, G and having node X_j pointing toward X_i. Mathematically, the joint distribution of all node values in DAG can be decomposed as the product of the local distribution of each node and its' parents:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i)) \qquad (2)$$

The Markov Blanket is another important characteristic of the Bayesian network. The Markov Blanket of a variable is the set of variables that completely shield off this variable from the other variables. To such an importance aim and for the purpose of classification, this study focuses on the Markov Blanket to identify minimal set of variables that are required to predict the metastasis outcome

3.1 Structure Learning of Bayesian Network

There are generally two main approaches to construct the Bayesian network from data: (1) the score and searching approach and (2) the dependency analysis approach. The first approach involves measuring fitness of structure and searching for the best structure that describes the data. Several scoring methods have been applied, including the Bayesian scoring, entropy based and minimum description length. Once the scores are obtained, the next step is to use search methods, such as the heuristic search, to build the best-fit network structure. Although this approach have been used by many researcher in reconstructing GRN, it mainly suffered from computational complex search and disability to provide posterior distributions over all the parameters of the model that are needed to quantify uncertainty in the gene regulators.

The dependency analysis approach or constraint-based learning on the other hands, aims to identify from the data the dependencies to construct the network structure. In this study, we proposed to low-order conditional independence and its variants, full-order conditional independence, to construct a cellular network.

Full-order conditional independence is the exact set of edges between the successive variables X_j and X_i, given the remaining variables X_{Vj}, $V_j = V \setminus \{j\}$ and $X_{Vj} = \{X_k;\ k \in V_j\}$. It can be defined as:

$$\tilde{G} = \left(X,\ \left\{(X_j, X_i);\ \cancel{X}_i \perp X_j \mid X_{V_j}\right\} i, j, \in V\right) \tag{3}$$

DAG $\tilde{G}$ is the smallest sub graph to which the probability distribution P has allows for a Bayesian network representation. Reverse discovery of DAG $\tilde{G}$ to model a cellular network requires determining each variable X_i and the set of variables X_j on which variable X_i is conditionally independent given the remaining variables X_{Vj}. Hence, by using the Equation 3, this approach has extended the principles of concentration graph that employed conditional independence to the Bayesian network case. However, by applying this approach, the curse of dimensionality is still a problem because the number of genes v is much greater than the number of measurements in n samples ($v >> n$) and conditional independence for each variables X_i given others remaining variable X_{Vj} is yet to be computed.

To reduce the high dimension of gene expression data, qth order conditional independence, DAGs $G^{(q)}$ (whereby $q < v$) is estimated from DAG $\tilde{G}$. By doing so, the Bayesian network is extended based on the consideration of low-order conditional independence, and this is similar to the work of Wille and Buhlmann [13] for the GGM. DAGs $G^{(q)}$ is defined as below:

$$\forall q < v, \ G^{(q)} = \left(X, \ \left\{ \left(X_j, X_i \right); \ \forall Q \subseteq V_j, |Q| = q, \cancel{X}_i \perp X_j \mid X_Q \right\} i, j \in V \right) \quad (4)$$

$$\forall \, q < v, \ \tilde{G} \subseteq G^{(q)} \quad (5)$$

DAGs $G^{(q)}$ is different from $\tilde{G}$ but it provides an alternative way of producing dependence relationship between variables, which is particularly suited for sparse network such as gene networks. However, DAGs $G^{(q)}$ is no longer associated with global relationship in the Bayesian network representation. Nevertheless, DAGs $G^{(q)}$ circumvents heavy statistical tasks and computation costly search in large number of variables. For additional technical details on this proposed method please refer to Ahmad et al. [6].

3.2 Revising Gene Regulatory Relationship with Integration of Transcription Factors

In a nutshell, two genes are regulating if transcription factor of regulator gene can bound at promoter region of target genes. Using such intrinsic biological feature, the regulatory relationships obtain by using the proposed Bayesian network model are verified. A number of necessary bioinformatics toolkits are wrapped to identify significant regulatory relationships. Three bioinformatics toolkits; (1) Ensembl, (2) TFSearch and (3) TRANSFAC, and their corresponding website are used in this study.

The name of dependent gene, T_h is entered in Ensembl and 1000 base pair upstream DNA sequence is then selected as a promoter region of target gene. This sequence is then copied and used as an input in TFSearch tool to find all possible transcription factors that can bind to a given promoter region. The regulator gene, T_g is examined using TRANFAC. This tool presents a list of transcription factors that are associated with regulator genes and DNA binding motif. If the transcription factor for both dependent gene and regulator gene match, then there is a dependency between X_i and X_j where $X_i \rightarrow X_j$.

3.3 Dataset Description

We tested this proposed method using a data set of 97 breast cancer microarray from van't Veer et al [14]. These cohorts of breast cancer patients are 55 years old or younger. We obtained this data from the Integrated Tumor Transcriptome Array and Clinical data Analysis database (ITTACA, 2006). Among the remaining 97 samples, 46 developed distant metastasis within 5 years and 51 remained metastasis free for at least 5 years. DNA microarray analysis was used by van't Veer to determine the expression levels of approximately 25,000 genes for each patient.

4 Experimental Results and Discussion

To obtain insights into the mechanism of gene regulation and how gene mutations act to turn on tumor development and metastasis progression in a cellular network context, the proposed method is executed on the breast cancer dataset producing a GRN as shown in Fig. 1.

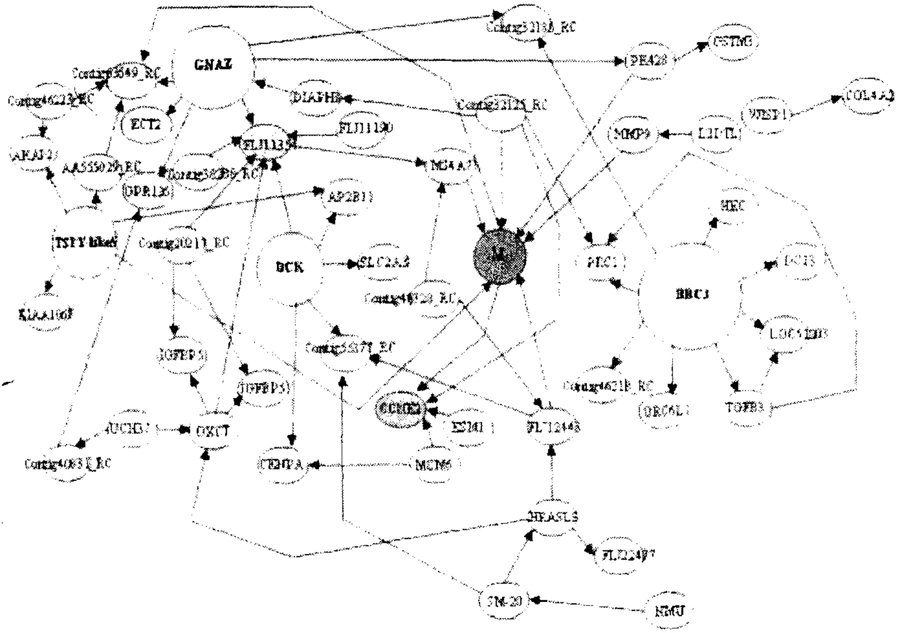

Fig. 1. The GRN for breast cancer metastasis using the low-order conditional independence method

This learned network revealed a group of genes which are primarily associated with causing metastasis, M. The larger nodes in the graph specify the genes when expressed at different levels lead to a major effect on the status of other genes (e.g., on or off). Meanwhile, the light-shaded nodes denote the highly regulated genes. Four genes that are found to regulate the expression levels of other genes are: BBC3, GNAZ, TSPY-like5 (TSPY5), and DCK. Two genes are highly regulated: FLJ11354 and CCNE2. This GRN involved 50 genes associated with metastasis, M, and 39 of them are annotated. Additionally, the p-value of the conditional independence test between the transcription regulatory genes and their co-expressed genes is given in [6].

Based on the experiment that has been carried out by using the proposed method, the relation between genes in the GRN is required to be verified. With the transcription factors of regulator gene and list of DNA binding sites of target gene at hand, as explained in Section 3.2, the regulatory relationship between G and H, can be examined. If the transcription factor of regulator gene TFg can bind to promoter site of target gene H, whereby $TF_h = TF_g$, then gene G and gene H could possibly has a relation. Together with this intrinsic biological features that play important role in the underlying regulatory mechanisms, 258 interactions are found to be biologically related. These interactions have fulfilled the biological test and hypothesis that are set earlier. Therefore, results as shown in Table 1 have been obtained. The results show that the proposed method works better with biological knowledge processing in comparison to network that rely on microarray only. In addition, Table 2 shows the p-values of some gene pairs (regulator and target genes) that involved in network inference model.

Table 1. Precision results for cellular network without/with biological knowledge processing. Both networks are constructed with 5000 genes.

Method	Total Edges	FP edges	TP edges	Precision %
Low-order conditional independence without biological knowledge	303	58	245	80.85
Low-order conditional independence with biological knowledge	258	43	215	83.33

In this experiment, transcription factors such as sp1, MyoD, GATA-1, GATA-2, GATA-3, CRE-BP; CREB, Ets, AP-1 or YY1 have been identified in nearly all of the interactions. Most of these transcription factors are discovered to be related to breast cancer metastasis, for instance, MyoD, AP-1 and sp1 have been identified by Mi *et al.* [15] to play important role in tumor progression, while Ets family of transcription factors are reported to be involved in cellular proliferation and apoptosis [16]. GATA transcription factor, particularly GATA-3 on the other hand has recently been identified as the key in controlling genes that involved in differentiation and proliferation of breast cancer [17]. Similar to the rest of transcription factors, CREB also has shown involvement in tumor initiation, progression and metastasis. It has been identified as proto-oncogene by Xiao *et al.* [18] and is found active in breast cancer, prostate cancer, lung cancer and leukemia cells. In addition, YY1 is discovered to play an essential role in tumorigenesis and is generally related to poor breast cancer prognosis [19].

5 Conclusion and Future Remarks

This paper described the need to integrate diverse data integration for better interpretation of GRN model. Two types of data integration approaches have been comprehensively explained; (1) homologous data integration and (2) heterogeneous data integration. Since most GRN models are mainly implemented based on microarray data, issues like reliability and quality concern are also debated by many researchers. The best available alternative is to integrate different data to address this problem and obtain a better understanding of the underlying gene regulatory mechanisms. Furthermore, with the currently available and enormous public databases, this effort appears to be the most promising since it utilizes the independent and complementary information to answer research questions.

The use of transcription factors to identify relevant regulatory interactions is the key idea in this research. In achieving this, three main bioinformatics toolkits for instance Ensembl, TFSearch and TRANFAC have been used. Each of these tools is used to apprehend the concept of biological intrinsic features of transcription factor and promoter. Based on the experiments that were conducted, 258 out of 303 interactions are identified to be biologically relevant. Furthermore, the p-value of regulated genes are computed to gain significant and efficient statistical results. The empirical results indicated several transcription factors such as sp1, MyoD, GATA-1, GATA-2, GATA-3, CRE-BP; CREB, Ets, AP-1 or YY1 play essential role in breast cancer metastasis. In the future, many more different data types will be integrated to obtain more insightful view of GRN and further facilitate our understanding of cancer growth.

References

1. Yavari, F., Towhidkhah, F., Gharibzadeh, S.: Gene regulatory network modeling using Bayesian networks and cross correlation. Biomedical Engineering Conference, CIBEC, Cairo (2008)
2. Gevaert, O., Van Vooren, S., Moor, B.D.: A framework for elucidating regulatory networks based on prior information and expression data. In: Eklund, P., Mann, G.A., Ellis, G. (eds.) ICCS 1996. LNCS, vol. 1115, pp. 240–248. Springer, Heidelberg (1996)
3. Huang, Z., Li, J., Su, H., Watts, G.S., Chen, H.: Large-scale regulatory network analysis from microarray data: Modified Bayesian network learning and association rule mining. Decision Support Systems 43, 1207–1225 (2007)
4. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using bayesian networks to analyze expression data. Journal of Computational Biology 7(3), 601–620 (2000)
5. Zou, M., Conzen, S.D.: A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics 21(1), 71–79 (2005)
6. Ahmad, F.K., Deris, S., Othman, N.H.: The Inference of Breast Cancer Metastasis through Gene Regulatory Networks. Journal of Biomedical Informatics (JBI) 45(2), 350–362 (2012)

Table 2. The prominent transcription factors for top ranked regulators

Regulator gene, G	Target gene, H	Transcription factors	P-value
BBC3	HEC	GATA-1; GATA-2; GATA-3; Oct-1;C/EBPalpha, C/EBPbeta	0.002178
BBC3	DC13	LyF-1; YY1; E2F; c-Ets-	0.001315
BBC3	Contig32185_RC	*	0.002261
BBC3	PRC1	c-Ets-; Oct-1; AP-1; c-Myb	0.002178
BBC3	ORC6L	AML-1a; USF-1; CP2; NF-Y	0.003732
BBC3	Contig46218_RC	*	0.021104
BBC3	LOC51203 (NUSAP1)	c-Ets-; LyF-1; HNF-3b; RORalp	0.003515
BBC3	TGFB3	E2F; AP-4; STATx; GATA-1; Sp1	0.013909
TGFB3	PRC1	SRY; c-ETS; GATA-1; SREBP-CRE-BP	3.60741E-06
TGFB3	LOC51203 (NUSAP1)	c-ETS; SRY; MyoD; GATA-1; AP-1	4.90607E-07
WISP1	COL4A2	AML-1a; Sp1; GATA-1; GATA-2; E2F;	0.595393
L2DTL	MMP9	GATA-1; GATA-3; NF-kap; AP-1; Sp1	3.53821E-05
PK428	GSTM3	GATA-1; SRY; HNF-3b; HNF-1; YY1	0.200820
GNAZ	PK428	USF; Sp1; MyoD; AP-2; YY1	0.007184
GNAZ	Contig32185_RC	*	0.020627
GNAZ	ECT2	Oct-1; LyF-1; MyoD; USF;	0.001813
GNAZ	Contig63649_RC	*	0.032360
GNAZ	FLJ11354	*	0.002381
GNAZ	GPR126	MyoD; CRE-BP; CREB; Oct-1; GATA-1	0.003125
TSPY-like 5	AP2B1	CREB; USF; GATA-1; GATA-2;GATA-3	0.042655

7. Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., et al.: Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proceedings of the National Academy of Sciences of the United States of America 101(25), 9309–9314 (2004)
8. Zhang, Y., Zha, H., Wang, J.Z., Chu, C.H.: Gene co-regulation vs. co-expression: The Pennsylvania State University, University Park, PA (2004)
9. Yeung, K.Y., Medvedovic, M., Bumgarner, R.E.: From co-expression to co-regulation: how many microarray experiments do we need? Genome Biology 5, R48 (2004)
10. Zhao, W., Serpedin, E., Dougherty, E.R.: Recovering genetic regulatory networks from chromatin immunoprecipitation and steady-state microarray data. Eurasip. J. Bioinform. Syst. Biol. (2008)
11. Kaleta, C., Göhler, A., Schuster, S., Jahreis, K., Guthke, R., Nikolajewa, S.: Integrative inference of gene-regulatory networks in Escherichia coli using information theoretic concepts and sequence analysis. BMC Systems Biology 4(116) (2010)
12. Pearl, J.: Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufmann Publishers Inc., Francisco (1988)
13. Wille, A., Buhlmann, P.: Low-order conditional independence graphs for inferring genetic networks. Statistical Applications in Genetics and Molecular Biology 4(32) (2006)
14. van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van de Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H.: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536 (2002)
15. Mi, Z., Guo, H., Wai, P.Y., Gao, C., Wei, J., Kuo, P.C.: Differential Osteopontin expression in phenotypically distinct subclones of murine breast cancer cells mediates metastatic behavior*. Journal of Biological Chemistry 279(45), 46659–46667 (2004)
16. Kato, T., Katabami, K., Takatsuki, H., Han, S.A., Takeuchi, K., Irimura, T., et al.: Characterization of the promoter for the mouse a3 integrin gene Involvement of the Ets-family of transcription factors in the promoter activity. Eur. J. Biochem. 269, 4524–4532 (2002)
17. Fang, S.H., Chen, Y., Weigel, R.J.: GATA-3 as a marker of hormone response in breast cancer. Journal of Surgical Research 157(2), 290–295 (2009)
18. Xiao, X., Li, B., Mitton, B., Ikeda, A., Sakamoto, K.: Targeting CREB for cancer therapy: friend or foe. Curr. Cancer Drug Targets 10(4), 384–391 (2010)
19. Gordon, S., Akopyan, G., Garban, H., Bonavida, B.: Transcription factor YY1: structure, function, and therapeutic implications in cancer biology. Oncogene 25, 1125–1142 (2006)