The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

Identifying differentially activated pathways by augmenting activities of transcription factor target genes

Hyunchul Jung[1] Eunjung Lee[1] Jongwon Kim[2]
Doheon Lee[1]

[1]Department of Bio and Brain Engineering, KAIST, 373-1 Guseong-dong, Yuseong-gu, Daejeon, 305-701, South Korea
[2]Department of Laboratory Medicine and Genetics, Sungkyunkwan University, School of Medicine, Samsung Medical Center, Seoul 135-710, South Korea

Abstract There are many approaches to discovering significant pathways in expression profiling. Our objective is to develop a robust approach to capturing differentially expressed pathways in disease expression profiling using pathways that have a transcription factor(TF). After collecting TF target gene, the TF target gene activity information in pathway and the pathway activity information are calculated using a t test. The p values from two analyses are integrated by the Chi-square inverse method. We evaluated our method using two cancer microarray data sets and comparing our results with those from two other pathway analysis methods. Our approach finds differentially expressed pathways that are directly associated with essential thrombocythemia, as well as general pathways that reflect cancer type-nonspecific traits and cancer type-specific pathways in breast cancer data. To test the robustness of our approach, we analyzed two lung data sets. Our approach also provides consistent and biologically reliable results. This study shows that it is better to find significant pathways in expression profiling where the TF target information is integrated within the pathway activity information. Our proposed approach is robust, and can also find cancer type-specific differentially expressed pathways.

1 Introduction

Translation of Genome-wide expression into biological meaningful data is still challenging. Much of the initial works have concentrated on the identification of differentially expressed genes and verification of their statistical significance in experiments. However, in most cases, biological insights can't be extracted from the identified differentially expressed genes because the interpretation of a lot of statistically significant genes is daunting job. To deal with this problem, recent efforts have interpreted microarray data by using prior knowledge such as ontologies and pathways. These researches aim at discovering biological pathways using genome-wide expression data.

Our approach considers the expression value of transcription factor (TF) targets for each pathway. Even though the use of the expression value of TF targets can be ambiguous, these expression values are strong signals directly measured from microarray, and

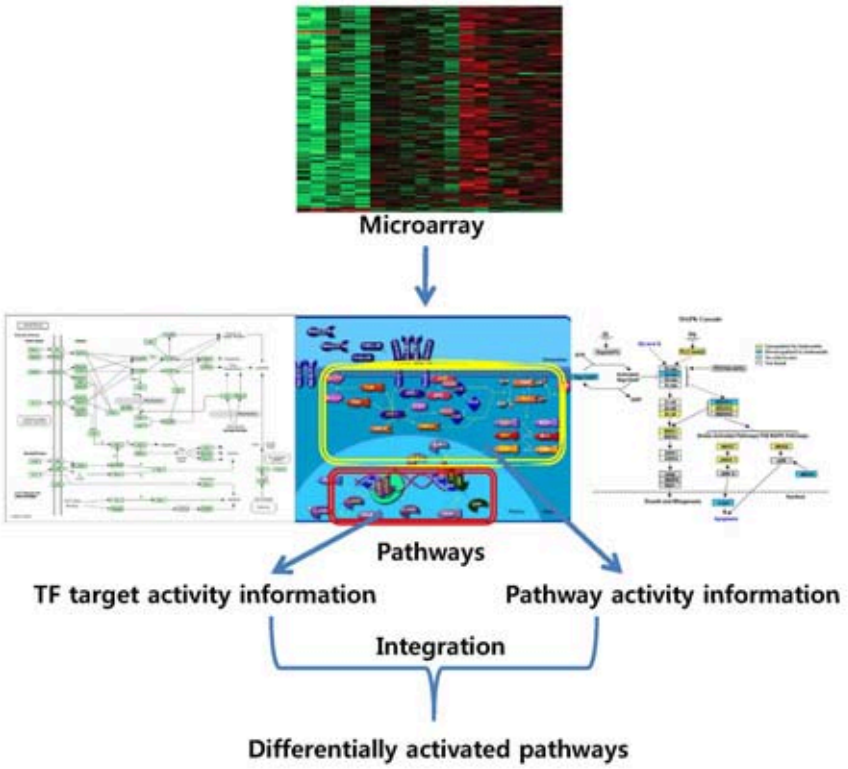

Figure 1: Outline of the methodology. Our approach finds differentially activated pathways by integrating the TF target genes expression information and the pathway genes expression information. Meta-analysis is applied to combine the p values from two different resources.

these values have correlations with irrespective pathway behaviors [5]. Therefore, the proposed method incorporates the pathway activity and the TF target activity information for each pathway (See Fig. 1). In terms of the pathway activity, we obtain the p value by Tian's method[3]. To obtain the TF target activity information, firstly we collect the TF list [6]. Among MsigDB pathways [7], we select canonical pathways having a TF. After that, we collect TF target genes from TRANSFAC database 11.0[8] and BZIP database [9]. Then we use the average of the TF target genes of t score for each TF. Lastly we compute the p value for each pathway from the null distribution constructed by permutation. The p value from both analyses is integrated through Chi-square inverse method [10]. This new proposed approach can find cancer type-specific perturbed pathways compared to GSEA and Tian's method. The advantages of the method are demonstrated by the cancer microarray data from essential thrombo-cythemia [11], breast cancer [12] and lung cancer [13,14].

2 Methods

2.1 Pathway and TF target information

Firstly we collected Human Transcription factor (TF) list (Messina et al,2004) In MsigDB, canonical pathway gene sets (c2.cp.v2.5.symbols.gmt) were used to select pathways which has at least one human TF in the list. Once we chosen 242 pathways, we collected TF target genes in TRANSFAC and BZIP database. Repressed targets by TF weren't considered, because repressed target information was limited compare to activated target information. Only activated targets by TF were considered in this analysis.

2.2 Data sets

[1] The essential thrombocyethemia data set [11] consists of samples from 16 patients, 9 of 16 have JAK2 V617F mutation and 7 of 16 patients do not have JAK2V617F mutation.

[2] The breast cancer data set [12] consists of samples from 286 patients, 209 of 286 patients are ER positive and 77 of 286 patients are ER negative.

[3] The Michigan group lung data set [13] consists of samples from 86 patients, 24 of 86 patients have poor outcome and 62 of 86 patients have good outcome. The Boston group lung data set [14] consists of samples from 62 patients, half of patients have poor outcome and the rest half of patients have good outcome.

2.3 p value

-The p value of pathway activities

Tian's method was used to infer pathway activities. Test statics for kth gene set can be written as T_k

$$T_k = \frac{1}{m_k} \sum_{i=1}^{m_k} t_i$$

Where m_k is the number of genes in the kth gene set and t_i is the t score of each gene in the kth gene set. The null distribution of $(T_1,\ldots,T_K)$ can be approximated by the empirical distribution of $(T_1^*,\ldots,T_K^*)$, where

$$T_k^* = \frac{1}{m_k} \sum_{i=1}^{m_k} t_i^*$$

t_i^* is permuted t_i. p value is calculated after 1000 permutations of t_i. Although Tian proposed normalized statistics NT_k, we used the p value of T_K(Not normalized) in this analysis.

-The p value of TF target activities

We use t test to detect location differences between two distributions. The test static for TF_k can be written as

$$TF_k = \frac{1}{M_k} \sum_{i=1}^{M_k} t_i$$

M_k represents the number of downstream targets of TF_k and t_i represents the t score of downstream target. After calculating each TF_k,

$$PTF_j = \frac{1}{j_N} \sum_{k=1}^{j_N} TF_k$$

PTF_j , the TF target activities of jth pathway, can be obtained by averaging TF_k in each pathway. $j_N$represents the number of TFs in jth pathway. For example, JAK2/STAT5 pathway has STAT5 and STAT3 TF. Firstly, the t value of STAT5 TF targets are added and divided by the number of STAT5 target genes. The same procedure is done with STAT3. Next, $PTF_{JAK2/STAT5}$can be obtained by averaging the two $TF_{k.}$. The p value is calculated through 1000 permutations of t_i.

2.4 Meta analysis

In order to combine the two p values from different sources, Meta analysis that is a set of classical statistical techniques to combine results from several studies was applied. Fisher's inverse Chi-square method is one of the meta analysis methods, and can be used to pool p values into a global p value. It consists in computing a combined statistic S from the different p values and

$$S = -2\log P_1 - \cdots - 2\log P_k$$

using this statistic for testing against a null hypothesis of a Chi-square distribution. The theoretical distribution of the summary statistic S under the null hypothesis is $\sim X^2_{2k}$.

In microarray, TF target activities are more directly measurable than other genes or proteins. Because of this fact, we gave weight(W) on the p value from TF target activities.

$$S = W(-2\log P_1) + (-2\log P_2)$$

P_1 is the p value from TF target activities and P_2 is the p value from pathway activities. Weight is given 2 on ET, Lung data and 3 on breast data. How to determine good weights given the data remains undecided at present.

3 Results

We have used this pathway analysis approach for several data sets. We evaluated our method with respect to showing valid results by comparing it to GSEA and Tian's method, using two cancer microarray data sets[10,11]. To test the roubustness of our method, we also analyzed data from two studies of lung cancer reported by the Boston group and Michigan group.

3.1 Myeloproliferative disorders – essential thrombocythemia(ET)

Myeloproliferative disorders (MPD) including polycythemia vera (PV), essential thrombocythe-mia (ET), and primary myelofibrosis (PMF) are characterized by a clonal expansions of a multipotent haematopoietic progenitor cell. ET is characterized of increasing bone marrow megakaryocytes, and persistent thrombocytosis[10]. Even though the existence of the JAK2V617F mutation has been reported in a high proportion of MPD

patients [15], only the 50% of the ET patients have this mutation. Regarding JAK2V617F positive ET patients, the constitutive kinase activity of JAK2V617F causes cytokine independent activation of JAK/STAT pathway, whereas JAK2V617F negative ET patients do not have activated JAK/STAT pathway [10]. Figure 2 shows a comparison among the GSEA, Tian's method and our approach. The most significant pathway in GSEA and Tian's method is G2PATHWAY. Other significant pathways including G2PATHWAY found by GSEA and Tian's method are cancer type-nonspecific pathways. These results do not give differentially expressed pathways linked to ET.

In contrast, our approach finds several other significant pathways that are directly associated ET. All pathways are directly associated with JAK2V617F mutation. For example, interleukin pathway families need JAK2 for signal transduction and as does the TPOPATHWAY, which is closely related with the production and differentiation of megakaryocytes[16]. GHPATHWAY describes the process by which growth hormone receptors dimerize on ligand binding and activate JAK2 protein kinase. In addition, Gleevec is effective in the treatment of essential thrombocythemia.

A. Enriched in ET patients with JAK2 mutation

Pathway name	P-value	Type
G2PATHWAY	0.0067	N
HSA04110_CELL_CYCLE	0.0311	N
HYPERTROPHY_MODEL	0.0363	N
CELL_CYCLE_KEGG	0.0448	N

B. ET patients with JAK2 mutation VS without JAK2 mutation

Pathway name	P-value	Type
G2PATHWAY	0.0211	N
PLK3PATHWAY	0.0488	N

P-value is < 0.05 according to NTK(Only upregulated pathway)

C. ET patients with JAK2 mutation VS without JAK2 mutation

Pathway name	P-value	Type
IL22BPPATHWAY	0.0027	S
TPOPATHWAY	0.0037	S
IL3PATHWAY	0.0116	S
HSA04630_JAK_STAT_SIGNALING_PATHWAY	0.0157	S
IL10PATHWAY	0.0202	S
BIOPEPTIDESPATHWAY	0.0249	S
GHPATHWAY	0.0333	S
GLEEVECPATHWAY	0.0387	S
IL7PATHWAY	0.0395	S
IL2RBPATHWAY	0.0455	S

Figure 2: A comparison between the results of GSEA (A) Tian's method (B) and our approach (C) for a set of genes associated with ET patients with JAK2 V617F mutation (p value < 0.05). The pathways marked by yellow (Type – N) represent cancer type-nonspecific pathways and the pathways marked by green (Type – S) show cancer type-specific pathways that is related with ET patients with JAK2 V617F mutation. Only our approach finds differentially expressed pathways that are related with essential thrombocythemia.

3.2 Breast cancer

Breast cancer patients who have estrogen receptors are said to be 'estrogen receptor positive', while those breast cancer patients who do not possess estrogen receptors are referred to as 'estrogen receptor negative'. These two types of breast cancer have different altered pathways and different treatment medications. Breast cancer patients who are estrogen receptor positive have cancer cell growth that is under the control of estrogen. However, it is said that estrogen receptor negative patients have different engines.

A. Enriched in ER positive patients

Pathway name	P value	Type
DREAMPATHWAY	0.003802	X
CARM_ERPATHWAY	0.021739	S
HSA04710_CIRCADIAN_RHYTHM	0.025243	X
OVARIAN_INFERTILITY_GENES	0.027132	X
BREAST_CANCER_ESTROGEN_SIGNALING	0.027778	S

B. ER positive VS ER negative

Pathway name	P value	Type
RNA_TRANSCRIPTION_REACTOME	0.0325	N
OVARIAN_INFERTILITY_GENES	0.0345	X
HSA04710_CIRCADIAN_RHYTHM	0.0491	X

P-value is < 0.05 according to NTK(Only upregulated pathway)

C. ER positive VS ER negative

Pathway name	P value	Type
CARM1PATHWAY	0.004409	S
VOBESITYPATHWAY	0.004705	S
HSA05030_AMYOTROPHIC_LATERAL_SCLEROSIS	0.011983	X
RBPATHWAY	0.014601	N
NUCLEAR_RECEPTORS	0.019436	S
ATRBRCAPATHWAY	0.018622	S
PLK3PATHWAY	0.018739	N

Figure 3: A comparison between the results of GSEA (A), Tian's method (B) and our approach (C) for a set of genes associated with breast cancer patients with estrogen receptor positive (p value < 0.05). The pathways marked by yellow (Type – N) represent cancer-type nonspecific pathways and the pathways marked by green (Type – S) show cancer type-specific differentially expressed pathways that are related with estrogen positive receptor breast cancer patients. Lastly the pathways marked by red (Type – X) do not have relationship with this disorder. Our approach finds more general cancer related pathways and differentially expressed pathways associated estrogen receptor positive breast cancer compared to two other methods.

We divided breast cancer data into estrogen receptor positive data and estrogen receptor negative data[12]. Figure 3 shows a comparison of differentially expressed pathways of estrogen receptor positive patients among the GSEA, Tian's method and our approach. GSEA can find not only cancer type-nonspecific pathways but also pathways closely associated with estrogen positive receptor breast cancer. CARM_ERPATHWAY predicted by GSEA describes the process whereby methyltransferase CARM1 methylates CBP and co-activates an estrogen receptor via Grip1. This pathway has a relationship with an estrogen receptor and BREAST_CANCER _ESTROGEN_SIGNALING that is directly related to this experiment. Tian's method finds only one cancer type-nonspecific pathway. The other

two pathways predicted by Tian's method seem not to have any relationship with breast cancer. On the other hand, our approach can detect both cancer type-nonspecific pathways and breast cancer related pathways. CARM1 is known to regulate estrogen-stimulated breast cancer growth[18]. NUCLEAR_RECEPTORS includes an estrogen receptor and ATRBRCAPATHWAY describes the phenomena connected with BRCA1 and BRCA2. Even though our approach finds one breast cancer unrelated pathway, it finds two cancer type-nonspecific pathways and four closely related pathways in the breast cancer estrogen receptor positive data.

3.3 Lung cancer

To test the robustness of our approach, we reanalyzed the lung cancer data that had been previously analyzed by GSEA. The aim of our approach, like that of GSEA, is not only to find differentially expressed tumor specific pathways but also to provide more consistent results than are obtained with single-gene analysis. GSEA reanalyzed data from two studies of lung cancer from the Boston group and the Michigan group. Even though GSEA found overlapping pathways in the two data sets, the results by GSEA were cancer type-nonspecific pathways, including cell cycle-related sets and p53 related sets. Figure 4 shows a comparison of commonly predicted differentially expressed pathways in both data sets among GSEA, Tian's method and our approach. GSEA only finds cancer type-nonspecific pathways and also pathways unrelated to lung cancer unrelated (pathways associated with breast and bladder cancer). Tian's method is more robust than GSEA. Even though it finds two pathways unrelated to lung cancer, it does uncover general cancer related pathways and lung cancer related pathways. Our approach also provides robust results, because it captures two angiogenesis related pathways in lung cancer. It is known that HIF-1α overexpression is a common event in lung cancer which may be related to the up-regulation of the angiogenic factor VEGF[19]. Our approach finds cancer type-nonspecific pathways as well.

4 Conclusion

This study shows that it is better to find significant pathways in expression profiling where the TF target information is integrated within the pathway activity information. Our proposed approach is able to find cancer type-specific differentially expressed pathways in the essential thrombocythemia data, whereas the other two methods cannot. In the breast cancer data set, our approach discovers more cancer type-nonspecific pathways and cancer type-specific pathways associated with breast cancer (estrogen receptor positive). Our approach also provides robust results with the lung cancer data sets. Mediating weight, adding more pathway sets and integrating with other pathway level analysis methods would uncover additional significantly differentially expressed pathways in a disease expression profile.

Acknowledgement

This work was supported by the Samsung Biomedical Research Institute Grant (C-A7-101-2) and by the Korean Systems Biology Program from the Ministry of Education, Science and Technology through the Korea Science and Engineering Foundation (No. M10309020000-03B5002-00000). In addition, this work was also supported by

A. Enriched in poor outcome

Pathway name	Type
HSA04110_CELL_CYCLE	N
HSA04115_P53_SIGNALING_PATHWAY	N
BREAST_CANCER_ESTROGEN_SIGNALING	X
HSA05219_BLADDER_CANCER	X

B. poor outcome VS good outcome

Pathway name	Type
P53HYPOXIAPATHWAY	N
BREAST_CANCER_ESTROGEN_SIGNALING	X
CIRCADIANPATHWAY	X
HSA04110_CELL_CYCLE	N
HSA04115_P53_SIGNALING_PATHWAY	N
HSA04510_FOCAL_ADHESION	S
VEGFPATHWAY	S

P-value is < 0.05 according to NTK(Only upregulated pathway)

C. poor outcome VS good outcome

Pathway name	Type
P53HYPOXIAPATHWAY	N
HSA05211_RENAL_CELL_CARCINOMA	X
VEGFPATHWAY	S
HIFPATHWAY	S
HSA04110_CELL_CYCLE	N

Figure 4: Overlapping pathways in two studies (p value < 0.05). A comparison is shown between the overlapping results of GSEA (A), Tian's method (B) and our approach (C) for a set of genes associated with the poor outcome of lung cancer in the Michigan and Boson data sets. The pathways marked by yellow (Type – N) represent cancer type-nonspecific pathways and the pathways marked by green (Type – S) show cancer type-specific differentially expressed pathways that are related with the poor outcome of lung cancer. Lastly the pathways marked by red (Type – X) do not have relationship with this disorder. Our approach provides consistent results with respect to cancer type-nonspecific pathways and pathways related to the poor outcome of lung cancer.

the second stage of the Brain Korea 21 Project in 2008. The authors also would like to acknowledge the support from the Korea Institute of Science and Technology Information (KISTI) Supercomputing Center and the support from the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Advancement)(IITA-2008-C1090-0801-0001). We would like to thank CHUNG Moon Soul Center for BioInformation and BioElectronics for providing research facilities.

References

[1] Khatri P, Draghici S, Ostermeier GC, Krawetz SA:Profiling gene expression using onto-express. Genomics. 2002 Feb;79(2):266-70.

[2] Khatri P, Draghici S:Ontological analysis of gene expression data: current tools, limitations,

and open problems. Bioinformatics. 2005 Sep 15;21(18):3587-95. Epub 2005 Jun 30.

[3] Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ.:Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci U S A. 2005 Sep 20;102(38):13544-9. Epub 2005 Sep 8.

[4] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP:Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50. Epub 2005 Sep 30.

[5] Breslin T, Krogh M, Peterson C, Troein C:Signal transduction pathway profiling of individual tumor samples.BMC Bioinformatics. 2005 Jun 29;6:163.

[6] Messina DN, Glasscock J, Gish W, Lovett M: An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. Genome Res. 2004 Oct;14(10B):2041-7.

[7] Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP: GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics. 2007 Dec 1;23(23):3251-3. Epub 2007 Jul 20.

[8] Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, Krull M, Matys V, Michael H, Ohnhauser R, Pruss M, Schacherer F, Thiele S, Urbach S: The TRANSFAC system on gene expression regulation. Nucleic Acids Res. 2001 Jan 1;29(1):281-3.

[9] Ryu T, Jung J, Lee S, Nam HJ, Hong SW, Yoo JW, Lee DK, Lee D. :bZIPDB: a database of regulatory information for human bZIP transcription factors. BMC Genomics. 2007 May 30;8:136.

[10] LV Hedges, I Olkin : Statistical methods for meta-analysis, Academic Press. 1985.

[11] Schwemmers S, Will B, Waller CF, Abdulkarim K, Johansson P, Andreasson B, Pahl HL: JAK2V617F-negative ET patients do not display constitutively active JAK/STAT signaling. Exp Hematol. 2007 Nov;35(11):1695-703. Epub 2007 Aug 30.

[12] Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA:Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005 Feb 19-25;365(9460):671-9.

[13] Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S:Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med. 2002 Aug;8(8):816-24. Epub 2002 Jul 15.

[14] Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M:Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A. 2001 Nov 20;98(24):13790-5. Epub 2001 Nov 13.

[15] James C, Ugo V, Le Couedic JP, Staerk J, Delhommeau F, Lacout C, Garcon L, Raslova H, Berger R, Bennaceur-Griscelli A, Villeval JL, Constantinescu SN, Casadevall N, Vainchenker W.: A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. Nature. 2005 Apr 28;434(7037):1144-8.

[16] Kaushansky K : Thrombopoietin the primary regulator of platelet production. Trends Endocrinol Metab. 1997 Mar;8(2):45-50.

[17] Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, Vogel V, Robidoux A, Dimitrov N, Atkins J, Daly M, Wieand S, Tan-Chiu E, Ford L, Wolmark N:Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. J Natl Cancer Inst. 1998 Sep 16;90(18):1371-88.

[18] Frietze S, Lupien M, Silver PA, Brown M:CARM1 regulates estrogen-stimulated breast cancer growth through up-regulation of E2F1.Cancer Res. 2008 Jan 1;68(1):301-6.

[19] Liu LZ, Fang J, Zhou Q, Hu X, Shi X, Jiang BH.:Apigenin inhibits expression of vascular endothelial growth factor and angiogenesis in human lung cancer cells: implication of chemoprevention of lung cancer.Mol Pharmacol. 2005 Sep;68(3):635-43. Epub 2005 Jun 9.