Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W619–W624***
doi:10.1093/nar/gkm469

DEEP—A tool for differential expression effector prediction

Jost Degenhardt, Martin Haubrock*, Jürgen Dönitz, Edgar Wingender and Torsten Crass

Department of Bioinformatics (Medical Faculty), Georg August University, Goldschmidtstraße 1, 37077 Göttingen, Germany

Received January 31, 2007; Revised April 13, 2007; Accepted May 29, 2007

ABSTRACT

High-throughput methods for measuring transcript abundance, like SAGE or microarrays, are widely used for determining differences in gene expression between different tissue types, dignities (normal/malignant) or time points. Further analysis of such data frequently aims at the identification of gene interaction networks that form the causal basis for the observed properties of the systems under examination. To this end, it is usually not sufficient to rely on the measured gene expression levels alone; rather, additional biological knowledge has to be taken into account in order to generate useful hypotheses about the molecular mechanism leading to the realization of a certain phenotype.

We present a method that combines gene expression data with biological expert knowledge on molecular interaction networks, as described by the TRANSPATH[1] database on signal transduction, to predict additional—and not necessarily differentially expressed—genes or gene products which might participate in processes specific for either of the examined tissues or conditions. In a first step, significance values for over-expression in tissue/condition A or B are assigned to all genes in the expression data set. Genes with a significance value exceeding a certain threshold are used as starting points for the reconstruction of a graph with signaling components as nodes and signaling events as edges. In a subsequent graph traversal process, again starting from the previously identified differentially expressed genes, all encountered nodes 'inherit' all their starting nodes' significance values. In a final step, the graph is visualized, the nodes being colored according to a weighted average of their inherited significance values. Each node's, or sub-network's, predominant color, ranging from green (significant for tissue/condition A) over yellow (not significant for either tissue/condition) to red (significant for tissue/condition B), thus gives an immediate visual clue on which molecules—differentially expressed or not—may play pivotal roles in the tissues or conditions under examination.

The described method has been implemented in Java as a client/server application and a web interface called DEEP (Differential Expression Effector Prediction). The client, which features an easy-to-use graphical interface, can freely be downloaded from the following URL: http://deep.bioinf.med.uni-goettingen.de

INTRODUCTION

It has always been a major goal of biological research to understand how the behavior of biological systems is governed by the properties of and interactions between their parts. Today's experimental high-throughput technologies allow us to simultaneously determine the state of thousands of system components at a cellular or molecular level. More and more powerful bioinformatics methods are applied to high-throughput data in order to unravel the interaction networks underlying a system's apparent behavior. Examining those networks in aberrant systems, e.g. malignant tissues, may lead to new insights into molecular etiology and thus suggest possible targets for the development of new therapeutic drugs.

A wide range of technologies is available for obtaining data on the expression level of genes. Most frequently used are methods like SAGE or microarrays, which measure transcript abundances as an estimate for gene expression height. In principle, both methods allow for two different

*To whom correspondence should be addressed. Tel: +49 551 3914915; Fax: +49 551 3914914; Email: martin.haubrock@bioinf.med.uni-goettingen.de

[1]TRANSPATH is a registered trademark of BIOBASE GmbH, Wolfenbüttel.

experimental set-ups (1,2): (i) Comparison of expression levels between different tissues or the same tissue under different conditions (treated/untreated) or dignities (normal/malignant). In this case, all genes which are over-expressed in one sample with regard to the other one are considered to be co-regulated and are generally believed to be also functionally related ('guilty by association'). (ii) Time series experiments typically monitor the temporal change in expression in just one tissue type after a stimulus has been applied. Genes for which changes in expression height are observed at an earlier time point are suspected to have a causal influence on the expression of genes displaying altered expression levels at a later stage. Various methods exist to derive possible gene interaction networks from such data.

In either case, the detection of significant expression differences may be improved by taking into account other sources of information, e.g. on the participation of the examined genes in common biochemical pathways (3). However, is has turned out not to be sufficient to rely on expression levels alone in order to generate useful hypotheses about the gene interaction networks eventually responsible for the realization of a certain phenotype. Rather, additional biological knowledge seems indispensable again for narrowing down the vast number of possible interactions that can be deduced from observed expression heights to a realistic and comprehensible amount. A resource frequently used for augmenting expression analysis methods is GeneOntology (4), which may provide additional hints on whether genes found to be co-expressed or inducing each other's expression do also share a common functional context. However, even for knowledge-based methods, additional (manual) expert evaluation is still indispensable in order to clarify which of the generated hypotheses on causal relationships seem plausible and should be subject to further experimental examination.

The majority of the available gene expression analysis methods takes primarily into account those genes which have been observed to be expressed differentially between the different conditions or time points; genes which are not found to be differentially expressed are considered to be of lesser interest. These approaches, however, neglect the fact that a gene doesn't necessarily have to be differentially expressed in order to exert effects specific for, e.g. a certain tissue or disease state. Rather, its gene product may just be subject to functional modulation by other, differentially expressed genes, rendering it functional under the one condition (where the modulator is over-expressed or, in case the modulator acts as an inhibitor, under-expressed) and inactive under the other (modulator under- or over-expressed, respectively). Thus by being dependent in its activity on a differentially expressed gene, such a non-differentially expressed gene acts as an effector specific for a certain expression profile. Consequently, if (i) a gene (or gene product) Y is known to be modulated in its activity by a gene X at least under some circumstances and (ii) X is found to be differentially expressed between the examined conditions A and B, we generate the hypothesis that Y's activity will also differ between conditions A and B according to X's abundance and hence may be called a (putative) expression-profile-specific effector.

In order to identify such additional effectors, one has to identify among non-differentially expressed candidates those genes (or gene products) which are already known to be subject to modulation by differentially expressed ones. We present a method (and provide an implementation thereof, called DEEP—Differential Expression Effector Prediction) to identify such molecules by applying already existing biological expert knowledge about biomolecular interaction networks, as provided by resources like the TRANSPATH database on signal transduction (5), to user-supplied, newly generated gene expression data. DEEP thus also demonstrates how the vast amount of information on potential interactions contained in databases like TRANSPATH can be utilized by, e.g. filtering for data sets of actual relevance in a certain expression background.

METHODS AND IMPLEMENTATION

Possible data sources and gene filtering

DEEP has initially been designed for the analysis of user-supplied SAGE data in the form of two sets (one for each condition A and B) of absolute tag counts N_i, mapped to their corresponding genes G_i. Alternatively, DEEP can access all human SAGE libraries as provided by CGAP on their SAGE Genie server (6), pooled according to CGAP's organ/tissue and dignity (normal/cancer/tumor associated) classification. If such a library set is selected, or more than two SAGE gene lists have been supplied for any of the two conditions, the corresponding list set is being merged into a 'meta library' by summing up the tag counts for each gene found in at least one of these lists. In any case, P_i values for F-fold over-expression (default: $F = 2$) are calculated using the method described in (1). Eventually, all P_i values are re-mapped to significance values s_i, with $s_i = 1 - P_i$ in case of over-expression in tissue/condition A and $s_i = P_i - 1$ otherwise. Only those genes with $|s_i|$ values exceeding a user-defined threshold are processed further.

Apart from direct support for SAGE data, DEEP is, in principle, capable of processing all kinds of expression data sets that consist of a list of genes G_i with corresponding significance values $-1 \leq s_i \leq 1$. Such values can, for instance, easily be derived from microarray data, rendering DEEP also applicable to this widely used experimental method for assessing differential gene expression.

Mapping over-expressed genes to network nodes

Differentially expressed genes surviving the filtering process described above are mapped to corresponding molecule entries in the TRANSPATH Professional database (version 7.1) on signal transduction, using UniGene identifiers as common denominator. Since TRANSPATH contains only manually curated information on signaling events for which experimental evidence is available, and since TRANSPATH comprises signaling components only, not all expression values will thus be

assignable to a TRANSPATH molecule; in fact, only ~12% of all tags found in the SAGE-Genie libraries could be successfully mapped to TRANSPATH. If, on the other hand, more than one SAGE tag is linked to one TRANSPATH molecule, their counts are averaged during the mapping process.

TRANSPATH features a molecule classification schema with various levels of abstraction (7); the mapping process described above, for instance, yields a list of so-called 'basic molecules', which represent species-specific sequence variants. For easing further processing, the identified basic molecules are further mapped to TRANSPATH 'ortholog' entries S_j, which subsume all species-specific variants. All S_j finally get their corresponding differentially expressed gene's s_i assigned to them.

Network reconstruction and effector prediction

The differentially expressed ortholog molecules S_j identified in the previous step are used as starting points for reconstructing a signal transduction network, represented as a graph with molecules as nodes N_k and signal transduction events (also taken from TRANSPATH) as directed edges. Reconstruction is performed by depth-first searching the network implicitly described by TRANSPATH to a user-defined depth, i.e. by calling the *buildNetwork* procedure depicted in the pseudocode outline (Figure 1) for each starting point S_j with parameters *buildNetwork*(j, 0).

In a subsequent step, the reconstructed network graph gets traversed, again starting from each node which represents a differentially expressed gene (or its corresponding TRANSPATH ortholog entries, respectively) S_j, by calling the procedure *propagateSignificance*(j, s_j, 0) from the pseudocode outline for each S_j). During this traversal, each encountered—and not necessarily differentially expressed!—successor node N_k 'inherits' the current starting node's s_j value. More precisely, each node keeps track of a weighted average a_k of the s_j values of all starting nodes it has been reached from, with $1/k_{jk}$ as weighting factor and k_{jk} being the number of steps that were required to reach N_k, starting from the node representing S_j. If an N_k gets visited several times—be it from the same S_j node via different paths, be it from different starting nodes—each visit will contribute to the node's a_k as described[2]. Consequently, each molecule node's a_k value is based on the significance by which one or more genes were considered differentially expressed in a certain experiment, yielding a measure for the degree to which its activity may be influenced in a tissue- or dignity-specific manner.

Finally, the graph is visualized, representing each molecule node's (initial or calculated) a_k value mapped to a color spectrum ranging from green ($a_k = 1$, i.e. specific for tissue/condition A) over yellow ($a_k = 0$, i.e. not tissue/condition-specific) to red ($a_k = -1$, i.e. specific for tissue/condition B). Thus a node's, or sub-network's, predominant color immediately gives a visual clue on which molecules—differentially expressed or not—may play pivotal roles in the tissues or conditions under examination.

```
max   : maximum search depth
g     : directed graph (initially empty)
a     : list of weighted averaged significance values
v     : list counting number of visits for each node
d     : list of minimum distances of nodes with regard to all starting nodes

procedure buildNetwork(k, depth)
    if g does not contain a node for molecule k then
        N ← new node representing k
        add N to g
        a[k] ← 0
        v[k] ← 0
        d[k] ← 0
    if depth ≤ d[k] and depth < max then
        d[k] ← depth
        for each interaction r having k on left-hand side do
            for each molecule l in right-hand side of r do
                createEdge(k, l)
                buildNetwork(l, depth+1)

procedure propagateSignificance(k, sStart, depth)
    v[k] ← v[k] + 1
    a[k] ← a[k] + sStart / (v[k] · (depth + 1))
    if (depth < max) then
        for each outgoing edge e of node N[k] do
            propagateSignificance(head(e), sStart, depth)
```

Figure 1. Pseudocode representation of the algorithms for reconstructing the signaling network and propagating the differentially expressed starting genes' significance values.

Treatment of different interaction types

Different types of interaction imply different semantics of the propagated s values. For instance, if an interaction between genes/molecules X and Y is known to be of type 'X inhibits Y', with X being over-expressed under condition A and Y not being expressed differentially, Y can be assumed to be more active under condition B than under condition A. This fact can be accounted for by inverting the propagated s value for inhibitory interactions. DEEP allows the user to choose which interaction types are to be treated this way and re-calculates the node coloring accordingly.

Calculating percolation clusters

In addition to reconstructing the signal transduction network in the above-described manner, DEEP also partitions the network into so-called signal percolation clusters. These clusters represent self-contained units of information flow, each starting from one differentially expressed gene and containing all its putative effector molecules found in the network. Percolation clusters hence represent the sub-networks, which may be subject to causal influence by a differentially expressed gene. Consequently, if another starting node S_l is encountered during the calculation of the starting node S_j's percolation cluster, S_l's cluster gets merged into the cluster currently under construction since it is well imaginable that S_l's activity may be modulated by S_j.

[2]Nodes representing starting points S_j just get their correspoinding s_j values assigned as a_j values.

Implementation

DEEP has been implemented in the Java programming language and consists of three parts: the DEEP core server, a JSP-based web interface and a downloadable Java client. This architecture prevents users from having to care for maintenance tasks, like updating the utilized databases. Just like the browser-based solution, the Java client communicates with the core server via firewall-friendly HTTP.

USAGE

Installation and usage

DEEP can be accessed in two ways. On the one hand, a web interface provides all basic functionality, while, on the other hand, a Java client is available which runs on the user's local computer. Both routes of access are available from the DEEP homepage (http://deep.bioinf.med.uni-goettingen.de). Since the local Java client allows for much more interactivity when inspecting the calculation results (e.g. zooming and panning), its use is strongly encouraged. It can be launched via a Java Web Start link, rendering its installation very straightforward; the only requirement is the presence of a moderately recent Java runtime environment (Java 5.0 or newer).

After launching DEEP either way, a new analysis can be initiated, and a 'wizard'-like interface will guide the user through the process of selecting the files containing the data to be analyzed or any of the pre-defined CGAP meta-libraries. Once network reconstruction is complete (which may take several minutes), the resulting network graph is presented; clicking on a node will display further information, including hyperlinks to corresponding external resources.

Examples

Hypotheses generated by tools like DEEP need to be verified either experimentally or by checking against previously published experiments. To this end, we present two literature-based case studies of DEEP usage.

Example 1. We compared three normal lung tissue data sets (Lung_normal_B_1, Lung_normal_CL_L15 and Lung_normal_CL_L16) from the pre-defined CGAP SAGE libraries and compared them with three corresponding malignant sets (Lung_adenocarcinoma-_B_1, Lung_adenocarcinoma_MD_L10 and Lung_adenocarcinoma_MD_L9) derived from lung adenocarcinoma samples, a subtype of nonsmall cell lung carcinoma (NSCLC). Network reconstruction was performed with $|s_i| \geq 0.9$ and search depth 2 (default values for all other parameters). As shown in Figure 2, interleukin 8 (IL-8) is one of the genes found to be over-expressed in malignant lung tissue and is hence used as a starting node for network reconstruction. Although not being differentially expressed themselves, the two G-protein coupled IL-8 receptors CXCR1 and CXCR2 are predicted as putative tumor-specific IL-8 effectors. In a second signaling step, various G-protein α subunits, including $G\alpha_i$, are also identified as putative targets of IL-8 in lung adenocarcinoma. All these predictions are in compliance with the corresponding literature, since apart from acting as a chemotactic and activating agent for leukocytes, IL-8 is known to support tumor growth in NSCLC by its angiogenic activity (8). Furthermore, IL-8 is suspected to serve as auto- and/or paracrine growth factor for NSCLC cells (9). The recent finding that the mitogenic effect of IL-8 is inhibited by pertussis toxin (10) suggest that signal transduction does indeed occurs under participation of G-protein α subunits belonging to the $G\alpha_i$ family. Obviously, all these observations could be confirmed using the DEEP tool.

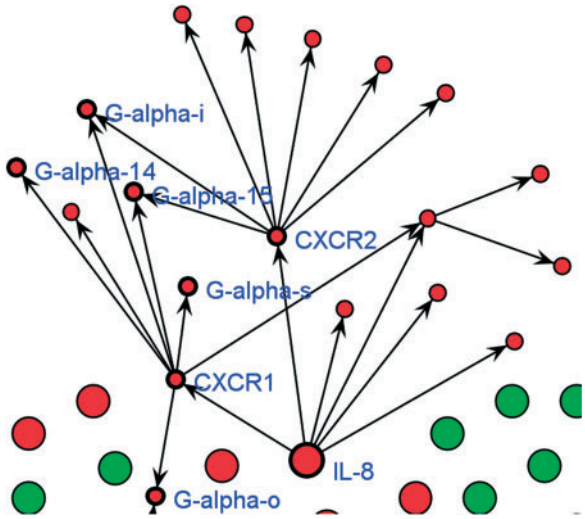

Figure 2. Reconstructed network using CGAP SAGE libraries derived from normal and malignant lung tissues (excerpt). Large nodes represent genes identified to be differentially expressed (green: normal tissue; red: malignant tissue), whereas small nodes stand for non-differentially expressed molecules, the coloring of which represents the degree to which they are predicted to act in an expression-background specific manner (green: active in normal tissue; red: active in malignant tissue; intermediate colors: specificity less clear; only effectors predicted to be tumor specific are present in the displayed part of the network). In concordance with experimental observation, DEEP correctly predicts the involvement of IL-8, its receptors CXCR1 and CXCR2 and various G-proteins in lung tumor development. This figure was created using the DEEP Java client.

A different line of NSCLC particularities could also be confirmed by DEEP. Fibronectin, an extracellular matrix glycoprotein, is known to be frequently over-expressed by NSCLC cells (11). It has furthermore been shown that fibronectin promotes NSCLC growth and metastasis by activating members of the AKT family of protein kinases, which normally participate in growth factor signal transduction (12). Apart from also identifying fibronectin as being over-expressed in lung cancer tissue, the analysis run described above also revealed AKT1 as possible fibronectin effector.

Example 2. The comparison of the whole collection of CGAP's normal mammary gland tissue libraries with their malignant counterparts provides another example for the suitability of DEEP for the interactive generation of hypotheses about expression-background specific effector genes. This analysis run has revealed kinase-associated

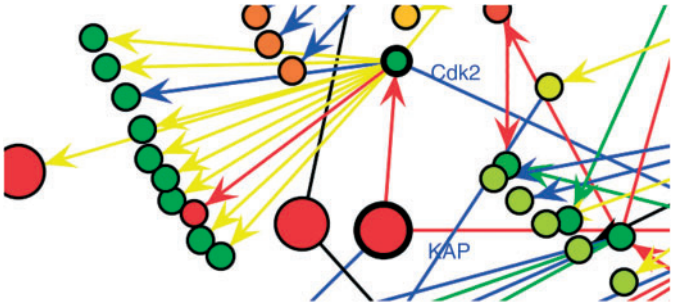

Figure 3. Reconstructed network using CGAP libraries derived from normal and malignant mammary gland tissue (excerpt), demonstrating DEEP's capability to handle the semantics of inhibitory interactions (red arrows): Since KAP is identified to be over-expressed in malignant tissue, DEEP predicts Cdk2 to be more active in the normal case. Other edge colors indicate activation (green), binding (blue), phosphorylation (yellow) *and other* (black). This figure was created using the DEEP Java client.

phosphatase (KAP) to be over-expressed in the malignant set (Figure 3), which complies to reports that KAP is over-expressed in various types of cancer, including breast cancer (13).

KAP participates in cell cycle regulation by interacting with cyclin-dependent kinases, like Cdk2, in an inhibitory manner. Since high levels of KAP expression should thus attenuate cell division, one would expect Cdk2 to be more active in normal tissue than in tumor cells, which seems to contradict the very nature of cancer. This puzzling conclusion is supported by DEEP if set up in a way to invert the sign of s_i values propagated through inhibitory interaction edges, as demonstrated in Figure 3. This apparent paradox has, in fact, only recently been resolved by the observation that aberrant splice variants of KAP (including a dominant-negative variant), which lead to increased Cdk2 activity, are expressed in some cancers (14). So albeit DEEP predicts the opposite of what has been observed regarding Cdk2 activity, this prediction was completely in line with what a human researcher would have concluded from the available knowledge about the involved molecules and their interactions.

These examples clearly demonstrate that DEEP is indeed capable of generating valuable hypotheses about genes which may act in a tissue- or condition-specific manner, though not being differentially expressed themselves.

OUTLOOK AND CONCLUSION

Future development

Although being already quite useful and usable in its current state, DEEP is, of course, still at the beginning of its evolution. For instance, a simple further plausibility check for predicted non-differentially expressed effector genes could be implemented by verifying that their transcripts have been found to be expressed at all in the original data; if not, the corresponding nodes should be omitted from the constructed graph. (In case of micro-array data, this criterion, of course, only holds if there has been any probe for the transcript in question on the chip.)

Another straightforward way of extending the system would be to utilize databases other than TRANSPATH, which contain data on biomolecular interaction networks. The next data resource projected to be thus included into the DEEP server is KEGG's LIGAND (15) section, providing information about the enzymes and metabolites involved in metabolic pathways. In a metabolic network, two enzyme genes will be considered to interact if one enzyme's product can serve as the other enzyme's substrate, or vice versa. The Reactome knowledgebase of biological pathways (16) is another candidate for inclusion.

Finally, the inclusion of molecules upstream to differentially expressed genes into the reconstructed network might also be worth implementing. A propagation of s_i-values 'backwards' through (directed) interaction edges will lead to the identification of molecules preceding differentially expressed gene products in, e.g., signaling cascades. The molecules thus identified can be considered as only being enabled to trigger certain functions in a tissue- or condition-specific manner by differentially expressed genes.

Conclusion

We have shown that by combining gene expression data with biological knowledge about biomolecular interaction networks, additional genes (or gene products) can be identified which may play distinct roles in different tissues or under different conditions, though not being differentially expressed themselves. While most gene expression analysis methods focus only on those genes found to be expressed differentially and hence will not present additional effectors in their result lists, our method extends the range of genes that may be crucial for the processes under examination, thus shedding new light on our understanding of the molecular basis of physiological processes as well as their pathological aberration, culminating in the prediction of new plausible targets for rational drug design.

ACKNOWLEDGEMENTS

This work was partially supported by 'Nationales Genomforschungsnetz' (NGFN, grant no. 01GR0480) and 'Intergenomics' (grant no. 031U210B), both German Ministry of Education and Research (BMBF). Funding to pay the Open Access publication charges for this article was provided by the Medical Faculty of Georg August University, Gättingen.

Conflict of interest statement. None declared.

REFERENCES

1. Steinhoff,C. and Vingron,M. (2006) Normalization and quantification of differential expression in gene expression microarrays. *Brief. Bioinformatics*, **7**, 166–177.
2. Lal,A., Lash,A.E., Altschul,S.F., Velculescu,V., Zhang,L., McLendon,R.E., Marra,M.A., Prange,C., Morin,P.J. *et al.* (1999) A public database for gene expression in human cancers. *Cancer Res.*, **59**, 5403–5407.

3. Subramanian,A., Tamaya,P., Mootha,V.K., Mukherjee,S., Ebert,B.K., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA.*, **102**, 15545–15550.
4. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
5. Krull,M., Pistor,S., Voss,N., Kel,A., Reuter,I., Kronenberg,D., Michael,H., Schwarzer,K., Potapov,A. *et al.* (2006) TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res.*, **34**, D546–D551.
6. Lash,A.E., Tolstoshev,C.M., Wagner,L., Schuler,G.D., Strausberg,R.L., Riggins,G.J. and Altschul,S.F. (2000) SAGEmap: a public gene expression resource. *Genome Res.*, **10**, 1051–1060.
7. Choi,C., Crass,T., Kel,A., Kel-Margoulis,O., Krull,M., Pistor,S., Potapov,A., Voss,N. and Wingender,E. (2004) Consistent re-modeling of signaling pathways and its implementation in the TRANSPATH database. *Genome Inform.*, **15**, 244–254.
8. Masuaya,D., Huang,C.-I., Liu,D., Kameyama,K., Hayashi,E., Yamauchi,A., Kobayashi,S., Haba,R. and Yokomise,H. (2001) The intratumoral expression of vascular endothelian growth factor and interleukin-8 associated with angiogenesis in nonsmall cell lung carcinoma patients. *Cancer*, **92**, 2628–2638.
9. Zu,Y.M, Webster,S.J., Flower,D. and Woll,P.J. (2004) Interleukin-8/CXCL8 is a growth factor for human lung cancer cells. *Br. J. Cancer*, **91**, 1970–1976.
10. Luppi,F., Longo,A.M., de Boer,W.I., Rabe,K.F. and Hiemstra,P.S. (2006) Interleukin-8 stimulates cell proliferation in non-small cell lung cancer through epidermal growth factor receptor transactivation. *Lung Cancer*, **56**, 25–33.
11. Han,J.-Y., Hong,S.K., Sug,H.L., Won,S.P., Jung,Y.L. and Nam,J.Y. (2003) Immunohistochemical expression of integrins and extracellular matrix proteins in non-small cell lung cancer: correlation with lymph node metastasis. *Lung Cancer*, **41**, 65–70.
12. Han,S.W., Khuri,F.R. and Roman,J. (2006) Fibronectin stimulates non-small cell lung carcinoma cell growth through activation of Akt/mammalian target of rapamycin/S6 kinase and inactivation of LKB1/AMP-activated protein kinase signal pathways. *Cancer Res.*, **66**, 315–323.
13. Lee,S.W., Reimer,C.L., Fang,L., Iruela-Arispe,M. and Aaronson,S.A. (2000) Overexpression of kinase-associated phosphatase (KAP) in breast and prostate cancer and inhbition of transformed phenotype by antisense KAP expression. *Mol. Cell. Biol.*, **20**, 1723–1732.
14. Yu,Y., Jiang,X., Schoch,B.S., Carroll,R.S., Black,P.M. and Johnson,M.D. (2007) Aberrant splicing of cyclin-dependent kinase-associated protein phosphatase KAP increases proliferation and migration in glioblastoma. *Cancer Res.*, **67**, 130–138.
15. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
16. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G.R. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**, D428–D432.