Mining to find the lipid interaction networks involved in Ovarian Cancers

Rajaraman Kanagasabai[1], Kothandaraman Narasimhan[3], Hong-Sang Low[2], Wee Tiong Ang[1], Aaron Z. Fernandis[2], Markus R. Wenk[2], Mahesh A. Choolani[3], and Christopher J. O. Baker[4*]

[1] Data Mining Dept. Institute for Infocomm Research, Agency for Science Technology and Research, Singapore,
[2] Department of Biochemistry, Faculty of Medicine, National University of Singapore, Singapore
[3] Department of Obstetrics and Gynecology, National University Health System (NUHS), Singapore.
[4] Department of Computer Science and Applied Statistics, University of New Brunswick, New Brunswick, Canada

ABSTRACT

The role of lipids in cancer during the genesis, progression and subsequent metastasis stages is increasingly discussed in the scientific literature. This information is discussed in a wide range of journals making it difficult for researchers to track the latest developments. A comprehensive assessment and translation of the *lipidome* of ovarian cancer, originating from literature, has yet to be made. We illustrate the deployment of semantic technologies; lipid ontology and text mining, in the aggregation and coordination of lipid literature. We provide the first report on the roles and types of lipids involved in ovarian cancer based on the mining of literature and identify key lipid-protein interactions that may point to potential drug discovery targets.

1 INTRODUCTION

Ovarian cancer (OC) is the fifth leading cause of all cancer-related deaths in women, and the most lethal of all gynaecological malignancies [1]. It is now becoming the most common gynecological cancer in developed countries. The disease presents late with only vague symptoms, and is one of the least understood cancers of all. If detected at an early stage, survival can be >95%, and in some cases chemotherapy can be avoided [2]. Lipids happen to play an integral part in the disease throughout the genesis, progression and subsequent metastasis stages [3, 4, 5]. Aberrant lipid synthesis can be detected during onset of ovarian cancer [6, 7, 8] and there are distinct associations between key regulatory proteins, onco-proteins and lipids [9, 10]. Apoptosis is considered one of the major events associated with ovarian cancer and rapid turnover of the cancer cell results in the release of tumour constituents into the blood stream. Lipids act as signaling molecules, initiators and mediators and play important roles during the cascade of apoptosis events. For example, ceramides impact signal transduction, and sphingosine-1-phosphate (S1P) and lysophospholipid are growth factors. There is immense value in the identification of biomarkers, which could be lipid related, and early detection techniques for ovarian cancer in serum and plasma samples. Identification of diagnostic biomarkers largely depends on the understanding of the complex interplay of biomolecules (protein and lipid) which are reported in the literature. A comprehensive assessment of the *lipidome* of ovarian cancer, originating from literature or unpublished experimental data, has yet to be made. Little is known about the distribution pattern or type and nature of lipids on different cell lines and cancer tissue samples of ovarian cancer. The complex network of lipids and proteins associated with ovarian cancer is inadequately understood.

In parallel a growing number of knowledge discovery systems are being developed which incorporate artificial intelligence (AI) techniques such as text mining and ontologies to deliver insights derived from the literature. [11,12,13]. The products of such techniques can take the form of automatically generated summaries, target sentences or lists of binary relations between entities from abstracts for which subsequent networks can be constructed and visualized in graphs, albeit without semantic labeling. The design of methodologies and tools to effect such information translation requires active input both from domain experts, knowledge and software engineers with specific attention to the scope of the translation task and application domain. Consequently the effective translation of information, derived using AI approaches, into actionable knowledge lags somewhat behind. In recent work [14, 15] we reported on the combination of semantic-based technologies; text mining, ontology population and knowledge representation, in the construction of a knowledgebase upon which we deployed data mining algorithms and visual query functionality. Integrated together these technologies can serve as a platform to translate information embedded in full-text scientific papers into semantically indexed aggregate of knowledge. We applied our ontology centric approach in the domain of lipidomics, addressing lipid-protein, protein-protein and lipid disease interactions with reference to apoptosis pathway discovery. This work was able to illustrate that our approach could accurately extract information and construct pathways equivalent to expert curation of documents. Moreover our visual query-answer paradigm was able to engage users in an interactive dialogue. In this next paper we again deploy the integrated text mining and semantics navigation infrastructure but this time use it to

* bakerc@unb.ca

provide insight into role of protein-lipid interactions in apoptosis pathways and ovarian cancer processes.

2 METHODS

The material for our analysis is scientific abstracts and full texts detailing studies into ovarian cancer in which we mine for proteins, lipids and interactions between these entities. The abstracts are obtained using our content acquisition engine that takes user keywords and retrieves the text and literature metadata by real-time crawling of Pubmed search results. In this case a collection of 7498 PubMed abstracts was identified by manual curation to be relevant to the subject of Ovarian Cancer (OC). Within this subset we found 683 papers that had lipid names from which 241 full papers were downloadable and mined for terms related to ovarian cancer, lipids, apoptosis and proteins. The motivation for selecting papers from this subset was manifold; we extended previous work [15] on lipids-protein interactions in apoptosis to ovarian cancer as key proteins such as PTEN, AKT, PI3K in apoptosis have been implicated to also play a role in the disease. Our combined expertise permitted us to evaluate the significance of text mining results in a coherent fashion. Retrieved research papers were converted from their original formats, to ascii text and made ready for mining by a customized document converter.

Our approach involves dictionary-based text mining and the indexing of sentences that mention pertinent interactions according to lipid ontology [14] scripted in the Ontology Web Language. The role of the ontology is to provide a query model to facilitate navigation of pertinent sentences by cancer researchers. On instantiation the ontology becomes a knowledgebase. Instances are generated from documents using with BioText. http://datam.i2r.a-star.edu.sg/~kanagasa/BioText/. Below we provide details of the tasks specific to knowledgebase population from ovarian cancer abstracts and selected full-text papers.

2.1 Knowledgebase Population

In its simplest form an OWL knowledgebase comprises of classes, instances of classes and properties (relations) linking instances. Instantiation of the knowledgebase comprises of three stages: Concept Instance Generation, Property Instance Generation, and Population of Instances. In the context of OWL-DL, Property Instances are assertions on individuals which are derived from relations found in predicate argument structures in mined sentences. In previous work we provided a performance evaluation of our instantiation pipeline [14, 15].

2.2.1 Concept Instance Generation. Concept instances are generated by first extracting the name entities from the texts and then normalizing and grounding them to the ontology concepts. Our entity recognizer uses a gazetteer that processes retrieved abstracts of documents and recognizes entities by matching term dictionaries against the tokens of processed text, and tags the terms found [15]. The lipid name dictionary was generated from Lipid Data Warehouse that contains lipid names from LIPIDMAPS [16], LipidBank and KEGG, IUPAC names, and optionally broad synonyms and exact synonyms. The manually curated UniProtKB/Swiss-Prot database[1] was used for creating the protein name dictionary. A disease name list was created from the Disease Ontology of Centre for Genetic Medicine (http://diseaseontology.sourceforge.net). We also constructed a list of hormones from UMLS[2]. A list of proteins and genes associated with ovarian cancer and apoptosis was manually created from Pubmed abstracts. A gold standard apoptosis pathway was constructed using proteins identified by manual consultation of pathway databases (KEGG for lipid binding proteins, Biocarta and IPA-Ingenuity Pathways Analysis) and subsequently cross checked manually with literature resources. The gold standard apoptosis pathway consists of 71 proteins (inclusive of isoforms and variants). This was further curated for subsequent incorporation into the Lipid Ontology. Our normalization and grounding strategy is as follows. Protein names were normalized to the canonical names entry in UniProtKB. Grounding is done via the UniProt ID. For lipid names, we define the LIPIDMAPS systematic name as the canonical name, and the LIPIDMAPS database ID is used for grounding. Hormone and Disease names are grounded via the UMLS ID. For apoptosis proteins, we augmented UniProt information, namely canonical Protein name, Alternative name, Gene name, Sequence Length, UniProt ID, GO: component, function and process.

2.2.2 Property Instance Generation. Object property and Datatype property instances are generated separately. From the Lipid, Protein, Hormone and Disease instances, ten types of relation pairs were mined, namely Protein (OC) – Protein (Apoptosis), Protein (OC) – Protein (OC), Protein (Apoptosis) – Protein (Apoptosis), Protein (Apoptosis) - Lipid, Protein (OC) – Lipid, Lipid – Lipid, Lipid - Hormone, Hormone-Hormone, Protein (OC) – Hormone, Protein (Apoptosis) – Hormone. For relation detection, we adopt a constraint-based association mining approach whereby two entities are said to be related if they co-occur in a sentence and satisfy a set of specified rules. The relation pairs from the resulting sentences are used to generate the Object property instances. Interaction sentences are instantiated as Datatype property instances. Several other Object

[1] http://www.uniprot.org/

[2] http://www.nlm.nih.gov/research/umls/

property instances are also generated to establish relations between, LIPIDMAPS Systematic Name and its associated IUPAC Name, synonyms and database IDs.

2.2.3 Population of Instances. We collect all the concept and property instances generated from the previous two steps to instantiate the ontology. Concept instances are instantiated to the respective ontology classes (as tagged by the entity recognizer), Object Property instances to the respective Object Properties and Datatype property instances to the respective Datatype properties. For this we wrote a script using the OWL programming framework, JENA API. http://jena.sourceforge.net/.

2.2 Knowledgebase Navigation

Navigation of the instantiated knowledgebase is achieved using Knowlegtor [14] visual-query navigation platform for OWL-DL ontologies which facilitates the construction of both complex concept level and keyword queries from OWL ontology constructs and relays them to an OWL reasoner to query a knowledgebase populated with A-box instances. The Knowlegator platform facilitates concept-instance, and binary-relation query as well as pathway discovery using graph mining algorithms. Graphical features of Knowlegator permit users to drag two proteins icons onto the query canvas and then invoke a search for transitive relations between these two entities [15].

3 LIPIDS: APOPTOSIS-OVARIAN CANCER

In this section we describe the results of our text mining analyses of abstracts and full-text papers and focus on the interactions between proteins and lipids, and characterization of these entities. We provide a narrative on the importance of these entities and interactions in the context of ovarian cancer, based on the research articles.

3.1 Interactions mined from the literature

Instantiated properties (relations between named entities) were enumerated according to the instances they linked and the type of interactions detected. Table 1 outlines the types and number of interactions in both bodies of text, abstracts and full text. Figure 1, shows the corresponding network representation of all interactions mined from the subset of 241 full-text articles. In both collections of documents interactions between proteins were clearly the most abundant whereas interactions between proteins and other named entities occurred at a lower level. Thus two data sets emerge, one representing well known established protein interactions and the other; a much larger number of interactions which may be novel, representing new leads, or false positives interactions. Despite the prevalence of protein-protein interactions we were able to identify only a small number of well known interactions between OC and apoptosis (AP) proteins in the abstracts collection. In this set only 53 of all interactions were mentioned more than 10 times, whereas a larger diversity of all interactions (1082) was mentioned 5 or fewer times. During curation of these interactions important yet unforeseen interactions between ovarian cancer proteins and apoptosis proteins were noted. In particular (i) OC_Bcl-x<-->AP_Bcl-2 interaction is relevant in apoptosis and involves mitochondria. (ii) OC_survivin<-->AP_Bcl-2: Survivin has an inverse association with Bcl-2 activity. Both are associated with early events in cancer and often reported to be found in endometrium. (iii) OC_EGFR<-->AP_p53 - Epithelial growth factor (EGFR) is a key player in epithelial ovarian cancer. Interaction of EGFR with p53 the master switch of apoptosis is an important finding.

Table 1. Interactions mined from the ovarian cancer bibliome. OC and AP represent a cancer and apoptosis pathway proteins respectively.

Interaction Type	**Abstract** (7498)	**Full Paper** (241)
OC-OC	223	13
OC-AP	505	195
OC-Lipid	11	14
OC-Hormone	8	1
AP-AP	113	59
AP-Lipid	10	8
Lipid-Lipid	3	23
Lipid Hormone	2	18
Protein Hormone	9	2
Hormone-Hormone	2	6

While these insights are poignant our primary interest focused on the lipidome of ovarian cancer and apoptosis and we sought to enumerate and characterize lipids and their biomolecular interactions. In summary the lipids we identified comprised predominantly of unsaturated fatty acids namely, species variants of octadecadienoic acid. Other categories of lipid were polyketides e.g. tacrolimus, and Isoprenoids. Most of the lipids that interact with hormones are under different subclasses of Isoprenoids, e.g. C15_isprenoids and C20 isoprenoids. Sphingolipids such as the ceramides were found, as were various sterol lipids such as ergostatetraen-3-ols. Some of the hormones were lipids e.g. estrogen related precursors.

3.2 Lipid-Protein Interactions

The role lipids play within apoptosis and ovarian cancer processes can best be elucidated through their ‘global’ interactions with biomolecules of known function as shown in Figure 1. Here we summarize a series of lipid-protein insights derived from the literature mining exercise and found to be relevant to our cancer experts. When human

neuroblastoma SH-SY5Y cells were exposed to in vitro oxidized polyunsaturated fatty acids, p53 was found to be significantly phosphorylated [17]. Subsequent screening of lipid peroxidation products in human neuroblastoma SH-SY5Y cells identified an oxidized lipid that potentially induces p53 phosphorylation. This lipid, 4-oxo-2-nonenal (ONE), is a recently identified aldehyde originating from the peroxidation of {omega}-6 polyunsaturated fatty acids. This evidence illustrates an important association of lipids, proteins and apoptosis and mining of full text documents was able to highlight a series of specific tumour suppressors as lipid interaction partners, see: http://datam.i2r.a-star.edu.sg/~kanagasa/stb09/TS

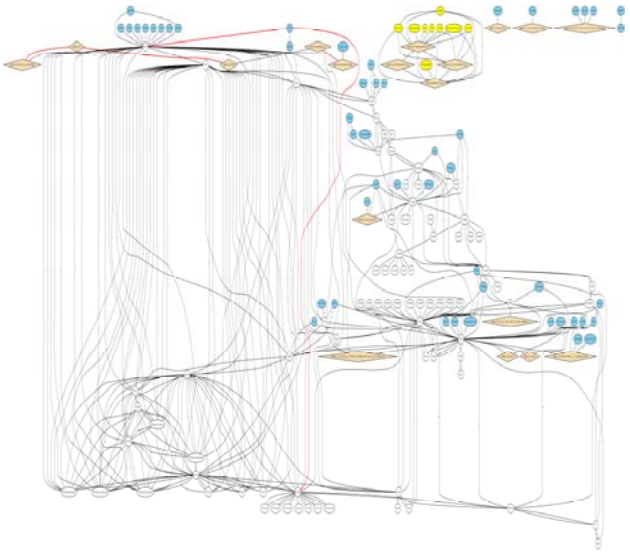

Fig. 1 Text mining-derived network of interactions between OC proteins, AP proteins, lipids and hormones. Ovals represent proteins. White are proteins on the apoptosis pathway, blue are proteins found in the OC abstracts, yellow represent hormones and diamonds represent lipids. Lines show the detected interactions and a numeric value represents frequency of detection of the interaction in the corpus. Lipid-tumor suppressor-apoptosis-cancer associations are red lines. The image is available online at http://datam.i2r.a-star.edu.sg/~kanagasa/stb09/Fig1

Three of the 12 tumor suppressors we identified (i) FHIT, (ii) RUNX3 and (iii) FANCF are linked to the lipid ergostatetraen-3-ol, a hydroxy vitamin D. The association of FHIT with ergostatetraen-3-ol appears to be a novel and useful insight. Expression of FHIT is reportedly associated with epithelial ovarian cancer [18] and more recently it was shown that FHIT-proteasome degradation is caused by mitogenic stimulation of the EGF receptor family in cancer cells [19]. Albeit indirect the ergostatetraen-3-ol lipid therefore has some relevance to OC. Similarly an indirect association between RUNX3 and lipids has been identified through text mining. RUNX3 is known to mediate apoptosis and cell growth inhibition in gastric epithelial cells. It is a candidate tumor suppressor that is frequently absent in gastric cancer cells. RUNX3 is rarely associated with ovarian cancer other than its involvement with CpG methylation of Granulosa cell tumours of ovarian origin [20]. A further lipid connection could provide new leads connecting RUNX3 with ovarian cancer especially ovarian tumours of epithelial origin. Thirdly a tumour suppressor FANCF is associated with ergostatetraen-3-ol. FANCF is a well studied gene for ovarian cancer. In particular its methylation contributes to the chemoselectivity of ovarian cancer [21]. Chemotherapeutic agents (polyketides) with lipids backbone structures like ergostatetraen-3-ol were found to be linked, through our text mining to ovarian cancer proteins. Further investigation of an association between FANCF and chemo selectivity is merited. Tumour suppressor, BRCA1, was also found to affect lipid synthesis through its interaction with acetyl-CoA carboxy carboxylase [10] during cancer.

The protein Akt (Protein Kinase B) plays a pivotal role in protein lipidome of ovarian cancer [22]. It is known to directly affect 2 biological pathways in ovarian cancer, namely the anti-apoptosis/cell survival pathway and cell metastasis pathway. Our results were able to illustrate Akt interaction either directly or indirectly with several lipids. In particular we identified lysophosphatidic acid (LPA) which is known to bind to LPA receptors proteins (LPAR1/LPAR2). This binding event results in a signaling cascade, inclusive of PI3K, that lead to the activation of Akt. The biosynthetic precursor of LPA, namely phosphatidic acid was also found in the graph of text mining results to be associated with Akt, as was the lipid Phorbol, a known inhibitor of LPA–LPAR binding. These lipid compounds may point to additional potential drug discovery targets other than the conventionally presumed PI3K [23]. Increasingly, fatty acids are now being recognized as ligands controlling cancer proliferation and cell death [24].

4 SEMANTIC QUERY

The instantiated Lipid Ontology (knowledge-base) was loaded to the Knowlegator platform [14] and made available to cancer specialists through the interactive query interface. Figure 2 shows the construction, by drag and drop from the search menu, of a graphical query and the corresponding sentences retrieved from different publications. This approach facilitates users who find it challenging to navigate large graphs like the one shown in Figure 1. Moreover it provides direct access to source material i.e. sentences and document provenance information.

DISCUSSION

Most of the lipids mined during the current study were either directly or indirectly associated with apoptosis proteins or pathways associated with ovarian cancer. Beyond the characterization of these lipids mined from abstracts and full text articles we sought to identify a series of novel insights that our cancer specialists were unaware of, such as the BRCA1-lipid association. We highlight the importance of octadecadienoic acids, polyketides and isopre-

noids as well as specific lipids, ergostatetraen-3-ol, lysophosphatidic acid, phosphatidic acid, and phorbol. While a clear statement on whether this is novel is somewhat fuzzy, as it depends on the domain expertise of the end user, we were able to make available both graph based visualization and ontology-centric visual query infrastructure so that our cancer scientists were able to construct meaningful hypotheses, queries and derive new insights. In particular the observation that lipids have direct interactions with p53 was considered novel, as was the identification of direct and indirect associations of the AKT pathways with lipids. The association with oxidized LDL (OxLDL) was found to increase the DNA binding activity of p53 [25]. Our approach provides bench scientists with two advantages. We provide fast access to information aggregated from multiple relevant sources and make this information available in both graphical and interactive ways so that scientists have the freedom to mine for new leads. The interactive query-answer methodology takes advantage of the OWL-DL lipid ontology, with explicit semantics, as a query model and use of Knowlegator leverages on description logics (A-box query).

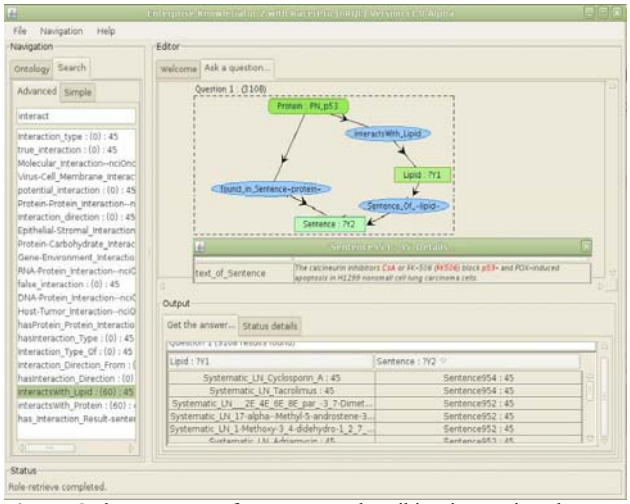

Figure. 2 shows a query for sentences describing interactions between lipids and p53. http://datam.i2r.a-star.edu.sg/~kanagasa/stb09/Fig2

ACKNOWLEDGEMENTS

Agency for Science & Technology Research, Singapore. National University of Singapore Office of Life Science (R-183-000-607-712), Academic Research Fund (R-183-000-160-112), Biomedical Research Council A*STAR (R-183-000-134-305), National Medical Research Council, Singapore (NMRC-07-066) awarded to M.R. Wenk. Bowen Li assisted with data processing.

REFERENCES

[1] Jemal A. etal. Cancer Statistics, 2008, CA Cancer J Clin 2008; 58:71-96

[2] Young RC. Mechanisms to improve chemotherapy effectiveness. Cancer. 1990 Feb 1;65(3 Suppl):815-22.

[3]Wang D, etal. S1P differentially regulates migration of human ovarian cancer and human ovarian surface epithelial cells. Mol Cancer Ther. 2008 Jul;7(7):1993-2002.

[4] Gupta GP, Massagué J. Platelets and metastasis revisited: a novel fatty link. J Clin Invest. 2004;114(12):1691-3.

[5] Wenk MR. The emerging field of lipidomics. Nat Rev Drug Discov. 2005;4(7):594-610.

[6] Fang X, etal. Lysophosphatidic acid is a bioactive mediator in ovarian cancer. Bichm Biophys Act. 2002; 1582(1-3): 257-64.

[7] Gercel-Taylor C, etal. Aberrations in normal systemic lipid metabolism in ovarian cancer patients. Gynecol Oncol. 1996 Jan;60(1):35-41

[8] Green AE, Rose PG. Pegylated liposomal doxorubicin in ovarian cancer. Int J Nanomedicine. 2006;1(3):229-39.

[9] Ahmed FE. Mining the oncoproteome and studying molecular interactions for biomarker development by 2DE, ChIP and SPR technologies. Expert Rev Proteomics 2008; 5(3):469-96.

[10] Moreau K, etal. BRCA1 affects lipid synthesis through its interaction with acetyl-CoA carboxylase, J Biol Chem 2006 Feb 10;281(6):3172-81

[11] Doms A, Schroeder M. GoPubMed: exploring pubmed with the geneontology. Nucleic Acids Res 2005 (33): W783–6.

[12] Müller HM,etal. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol 2004;2(11): 1984–98.

[13] Novácek V, etal., Infrastructure for dynamic knowledge integration—Automated biomedical ontology extension using textual resources, J. Biomed. Inf., Vol. 41, Iss. 5, pp 816-828

[14] Baker CJO ,etal. (2008). Towards ontology-driven navigation of the lipid *bibliosphere* BMC Bioinf. 2008, 9 (Supp 1):S5

[15] Kanagasabai R., etal., Ontology-Centric navigation of pathway information mined from text, 11th Ann.Bio-Ontologies Meeting, co-loc.ISMB 2008, Toronto Canada, July 20th 2008.

[16] Sud M., etal.LMSD: LIPID MAPS structure database. Nucleic Acids Res 2007, 35:D527-D532.

[17] Shibata T,etal. Identification of a Lipid Peroxidation Product as a Potential Trigger of the p53 Pathway. J. Biol. Chem. 281 (2): 1196-1204

[18] Ozaki K, etal. Impaired FHIT expression characterizes serous ovarian carcinoma. Br J Cancer. 2001;85(2):247-54.

[19] Bianchi F, etal. FHIT-proteasome degradation caused by mitogenic stimulation of the EGF receptor family in cancer cells. Proc Natl Acad Sci U S A. 2006;103(50):18981-6.

[20] Dhillon VS, etal. CpG methylation of the FHIT, FANCF, cyclin-D2, BRCA2 and RUNX3 genes in Granulosa cell tumors (GCTs) of ovarian origin. Mol Cancer. 2004;3:33.

[21] Olopade OI, Wei M. FANCF methylation contributes to chemoselectivity in ovarian cancer. Cancer Cell. 2003;3(5):417-20.

[22] Mabuchi S, etal. Inhibition of phosphorylation of BAD and Raf-1 by Akt sensitizes human ovarian cancer cells to paclitaxel, J Biol Chem 2002 Sep 6;277(36):33490-500

[23] Elstner E, etal. Ligands for peroxisome proliferator-activated receptorγ and retinoic acid receptor inhibit growth and induce apoptosis of human breast cancer cells *in vitro* and in BNX mice, PNA**S**, July 21, 1998; 95(15): 8806 - 8811.

[24] Hennessy BT etal. Exploiting the PI3K/AKT pathway for cancer drug discovery. Nat Rev Drug Discov. 2005 Dec;4(12):988-1004

[25] Mazière C, etal. Oxidized LDL induces an oxidative stress and activates the tumor suppressor p53 in MRC5 human fibroblasts. Biochem Biophys Res Commun. 2000 Sep 24;276(2):718-23.