Send Orders for Reprints to reprints@benthamscience.net

Protein & Peptide Letters, **2014,** *21,* **0000-0000**

Biomedical Hypothesis Generation by Text Mining and Gene Prioritization

Ingrid Petrič[1,2]*, Balázs Ligeti[3], Balázs Győrffy[4] and Sándor Pongor[2,3]

[1]*Centre for Systems and Information Technologies, University of Nova Gorica, Vipavska 13, SI-5000 Nova Gorica, Slovenia;* [2]*Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, Padriciano 99, I-34012 Trieste, Italy;* [3]*Faculty of Information Technology, Pázmány Péter Catholic University, Práter utca 50/A, H-1083 Budapest, Hungary;* [4]*Research Laboratory of Pediatrics and Nephrology, Hungarian Academy of Sciences, Bókay u. 53-54, H-1083 Budapest, Hungary*

Abstract: Text mining methods can facilitate the generation of biomedical hypotheses by suggesting novel associations between diseases and genes. Previously, we developed a rare-term model called RaJoLink (Petric *et al*, J. Biomed. Inform. 42(2): 219-227, 2009) in which hypotheses are formulated on the basis of terms rarely associated with a target domain. Since many current medical hypotheses are formulated in terms of molecular entities and molecular mechanisms, here we extend the methodology to proteins and genes, using a standardized vocabulary as well as a gene/protein network model. The proposed enhanced RaJoLink rare-term model combines text mining and gene prioritization approaches. Its utility is illustrated by finding known as well as potential gene-disease associations in ovarian cancer using MEDLINE abstracts and the STRING database.

Keywords: Biomedical hypothesis generation, text mining, disease gene prediction, gene prioritization, ovarian cancer.

1. INTRODUCTION

Research in life sciences is only possible today with access to online literature databases. This body of information is constantly broadening in scope, which presents a challenge to text mining researchers seeking to extract information for life scientists [1]. Hypothesis generation is a specific task in this large area. The term refers to generating a surprising or unexpected supposition based on information extracted from text resources [2]. From a data-mining perspective, text-based hypothesis generation is a case of link discovery [3], i.e. a hypothesis can be considered as a potential link between pre-existing knowledge items. In the genomics era, hypotheses are often formulated as relations involving molecular entities, such as genes, proteins, drugs, metabolites, etc. Text-based hypothesis generation approaches have been effectively applied in the detection of disease-drug interactions [4], gene-disease relationships [5], protein-protein interactions [6], pathway related information [7] as well as other bio-molecular events [8].

Previously we developed a computing methodology called RaJoLink which is designed to find associations between terms that are rarely used within a given domain of interest [9]. In this approach, a hypothesis is defined as an association between a rare, potentially interesting term and another term, such as a disease. In the present work we seek to extend this approach to more efficiently generate hypotheses relating to associations between diseases and genes. From a text-mining perspective, genes are terms defined in well-curated nomenclatures such as the HUGO Gene Nomenclature [10], so the task of incorporating them into a hypothesis generation framework would appear straightforward at first.

However, we have found that scientific articles use a variety of names for the same gene. This makes it difficult to differentiate between known and potentially novel gene-to-disease associations. For this reason we have explored the possibility of using an additional knowledge source, namely gene-to-gene association networks given in molecular databases such as STRING [11] or other protein-interaction databases to extend our methodology to find genes potentially associated with a given disease.

We disregard genes that are explicitly known to be associated with a target disease to identify genes that may be worth further examination in the future. We benchmark the method on marker genes in ovarian cancer [12], and show that prior biomedical literature can point towards genes subsequently shown to be linked to ovarian cancer. We introduce the background to literature-based hypothesis discovery and gene-disease associations (section 2) and describe the data and the computational methods used in this study (section 3). We then present the new methodologies for finding gene-disease associations and the application of these methods for discovering genes related to ovarian cancer (section 4). Sections 5 and 6 contain the discussions and conclusions, respectively.

*Address correspondence to this author at the Centre for Systems and Information Technologies, University of Nova Gorica, Vipavska 13, SI-5000 Nova Gorica, Slovenia, Tel: +38 65 3315231; Fax: +38 65 3315240;
E-mail: Ingrid.Petric@ung.si

0929-8665/14 $58.00+.00

2. BACKGROUND AND RELATED STUDIES

2.1. Swanson's Model

Swanson [13] first showed that bibliographic databases can serve as rich sources of undiscovered relations between already-existing data. Subsequently a powerful text analysis method was proposed and developed for generating hypotheses from previously disjointed sets of literature [14]. For many complex scientific problems, cooperation across the boundaries of different disciplines is often required. In such situations, software tools can support researchers in crossing domain boundaries by assembling separate pieces of knowledge into a logical context. This is extremely important for open knowledge discovery (hypothesis generation) where hypothesis target concepts are not yet defined [15]. On the contrary, the hypothesis of a closed discovery process is known, and the goal is to validate the connection between the hypothetically associated phenomena. In an open discovery process, only the investigating phenomenon is specified at the start of the process. If we are investigating a concept denoted with the term *a*, the open knowledge discovery process starts with having only the term *a* and the corresponding set of articles in which term *a* appears (called also literature *A*). Intermediate *B* terms are then used to discover the target term *c* (and the corresponding literature *C*). Term *c* is not defined *a priori*.

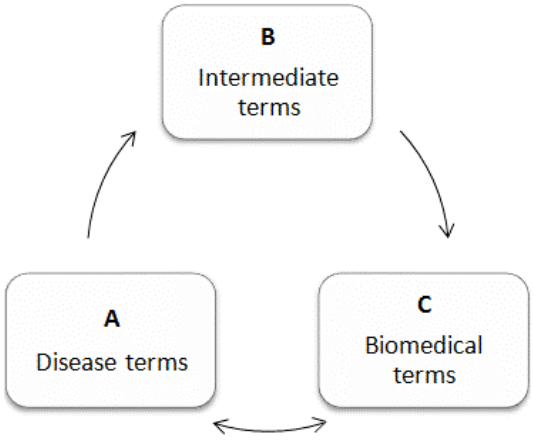

Figure 1. Swanson's ABC model for hypothesis generation [13,16].

Swanson's ABC model approach investigates whether an agent *A* is connected with a phenomenon *C* by discovering complementary structures via interconnecting phenomena *B* (as shown in Fig. **1**). If one literature discusses the relations between *A* and *B*, while a disparate literature investigates the relations between *B* and *C*, then the two literatures are complementary. The combinations of previously unknown but meaningful relations between *A* and *C* can be viewed as a new piece of knowledge that explains the phenomenon *A*. Consequently, following the complementary but disjointed literature items A-B on one hand and B-C on the other, one can hypothesize that there is a novel, plausible indirect association A-C if there are published relations A-B and B-C but no known direct relation A-C. From here we will use uppercase symbols *A*, *B*, and *C* to represent sets of terms (e.g., literature or collection of records), while lowercase symbols *a*, *b*, and *c* for single terms.

By considering unconnected sets of articles several new discoveries were made by Swanson. In one study, he discovered that Raynaud's syndrome can be treated with dietary fish oil [13]. Similarly, while studying the literature on migraines and separately the literature on magnesium, Swanson found implicit connections that were unnoticed at the outset of his research [16]. He noticed the possible relationship between the disjoint literatures by linking terms found in both sets of literatures. Prior to the Swanson's discovery, few researchers had paid attention to a direct migraine - magnesium or Raynaud's syndrome – fish oil connection. Numerous laboratory and clinical investigations started only after the publication of Swanson's insights. In particular, the migraine - magnesium example has become a gold standard in the medical literature mining field and has been used as a point of reference in several studies [2], [15], [17],[18].

2.2. Rare-term Model

To support literature-based link discovery, we developed a text mining model called RaJoLink [19]. This model examines rare pieces of information identified in biomedical text corpora and generates scientific hypotheses in a semi-automated way.

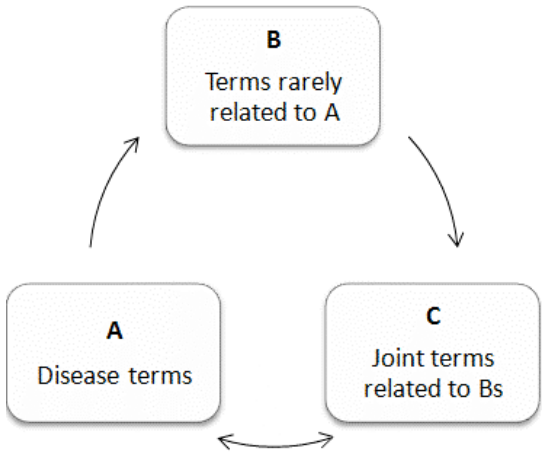

Figure 2. Hypothesis generation according to the original rare-term model.

In the original model the choice of candidate hypothesis was based on terms that rarely appear in the literature (see Fig. **2**). The method consists of three principal steps, *Ra*, *Jo* and *Link*, named after the key elements Rare terms, Joint terms and Linking terms, respectively. In the first step, Ra, the terms that rarely appear in literature about the phenomenon *A* are identified. In the second step Jo, MEDLINE articles (titles or abstracts) about the selected rare terms are inspected and joint terms that appear in the intersection of the literatures about rare terms are identified as the candidates for *C*. In order to provide explanation for hypotheses generated in step Jo, in the Link step the method searches for *b*-terms that connect the literature on joint term *c* (*c*-term) and

the starting literature *A*. In other words, steps Ra and Jo implement an open discovery process, while step Link corresponds to a closed discovery process, where *A* and *C* are both already known and the method is searching for *b*-terms. Methodological details of RaJoLink and its application to autism to generate new hypotheses are described elsewhere [19],[20].

2.3. The MeSH Classification

MeSH is a controlled vocabulary of the United States National Library of Medicine (NLM) created for indexing texts in the area of life sciences and to serve as a thesaurus that facilitates searching [21]. In the original rare-term model, the MeSH classification was used to map the terms from free text to the concepts within this controlled biomedical vocabulary. Each string variant that results from the text pre-processing and the morphological analysis of words was evaluated against the MeSH thesaurus. This way terms can be filtered out by selecting only those categories from the MeSH tree structure that the user is interested in. For example, the MeSH categories D12: Amino acids, peptides and proteins and D08: Enzymes and coenzymes are commonly selected if the user is interested in gene-disease associations.

2.4. The HUGO Gene Nomenclature

The HUGO Gene Nomenclature provides a unique and meaningful name as well as an abbreviation (gene symbol) for every known human gene [10]. A custom downloads interface of the HGNC database is publicly available at http://www.genenames.org/cgi-bin/hgnc_downloads.cgi.
The database stores data from individual researchers as well as from large-scale projects, such as the Human Genome Sequencing Consortium [22]. In the enhanced rare-term model, the HUGO gene nomenclature was used to extract gene symbols from MEDLINE abstracts.

2.5. The STRING Database

The STRING database [23],[24] is a precomputed resource for exploring associations between proteins and their genes. Functional associations are inferred from genomic associations, from functional descriptions, protein-protein interactions, and different evidence types integrated into confidence scores for prediction. The resource contains information for 1100 completely sequenced organisms and accessory information such as protein domains and 3D structures are also provided; the resource can be reached at http://string-db.org. We used the STRING database for re-ranking the results.

2.6. Finding Gene-disease Associations by Text Mining

Text mining has been successfully applied in finding various gene-disease associations [25], such as suggesting disease marker genes from MEDLINE records and ranking (prioritizing) genes based on biomedical literature [26]. Reviews of the earlier work are found in [27] and [28]. More recently, Hristovski and associates combined DNA microarray data and semantic relations extracted from MEDLINE, for generating novel hypotheses about potential drug therapies for Parkinson disease [29]. Frijters and colleagues also presented an application of their literature mining method in an open-ended retrieval of hidden relations for hypotheses in terms of gene-disease, drug-disease and drug-biological process associations in their studies of Graves' disease, milnacipran, pitavastatin and drugs that could interfere with cell proliferation [30].

2.7. Gene-disease Associations from Experimental Datasets

Traditionally, gene-disease associations are based on experimental data that have been evaluated manually. High-throughput techniques can now also mean that it is possible to experimentally compare the behavior of all human genes in healthy and diseased states. However, the evaluation of such lists is not simple [31]. Computational methods of "gene prioritization" were developed for this purpose [32]. Most of the methods combine the new experimental data with a background database containing information on co-occurrence, functional annotations, protein- protein interactions, pathways, and gene expression. Briefly, we can view new experimental data as numerical scores assigned to genes, and the background database as a network of genes in which the links are defined by one of the methods mentioned above. In the process of gene prioritization, the experimental scores are updated using the gene network data and the genes are re-ranked based on the new scores. Updating of scores can be based on graph distance (shortest path), on a propagation algorithm such as the popular PageRank [33] or on diffusion kernel methods [34], for example. The resulting methods differ in the kind of score updating methodology, the background database used, and most importantly, the size of the data they can handle. Relatively few methods can select genes from entire genomes or accept input data on all genes. For instance, it is customary to restrict the scope of candidate genes to a small region of the chromosome using methods of linkage analysis or to use known disease genes as a training set. One of our goals is to use approaches analogous to the methods of gene prioritization in order to further increase the sensitivity of hypothesis generation.

3. MATERIALS AND METHODS

3.1. Corpus Collection and Benchmark Datasets

The document sets in our experiments were acquired from the MEDLINE database through its PubMed system [35] using the Entrez Programming Utilities [36]. Each document set consisted of citations that comprised of abstracts obtained from PubMed by executing Boolean queries. The target sets of texts were restricted to abstracts of articles, because unlike the majority of full texts, they are freely available online in XML format.

The benchmark datasets were designed to test whether or not a method could efficiently predict that a gene plays a certain role, which was then experimentally confirmed later. For this test, we needed a corpus of abstracts published before a certain biological role was confirmed. We chose ovarian cancer as the model disease and used a recently published list of 37 ovarian cancer biomarker (OC biomarkers) genes [12] (Table **1**) as test cases. We then wanted to determine if the relationship between these genes and ovarian cancer could have been predicted on the basis of literature

published beforehand. In order to have a sufficient number of genes in the analysis, we selected the year 2007 as a separating line. A total of 10 OC biomarkers have been proposed after this date.

OC biomarker abstracts were selected using the search phrase: (biomarker OR biomarkers OR marker OR markers) AND (“cancer of ovary" OR "ovary cancer" OR "cancer of the ovary" OR "ovarian cancer" OR "malignant neoplasm of ovary" OR "malignant ovarian neoplasm" OR "malignant tumor of ovary" OR "malignant tumor of the ovary" OR "malignant neoplasm of the ovary" OR "malignant ovarian tumor" OR "malignant tumour of ovary" OR "ovarian malignancy" OR "ovarian carcinoma"). This search resulted in 4,878 abstracts published before the year 2007. We defined this set as the OC biomarker test corpus. Separately, 26,979 abstracts about the known OC biomarker genes [12] (Table **1**) published up until May 14th 2012 were obtained and these formed the OC biomarker prediction corpus. The data are deposited in supplementary datasets 1-2. We used the HGNC gene symbols, names and their synonyms (downloaded on December 23rd 2011). Such HGNC nomenclature was then applied to the terms that we automatically extracted from collections of MEDLINE abstracts.

3.2. Re-ranking Methods

From the mathematical point of view, genes selected by text mining analysis can be viewed either as an unranked set of gene names or as a ranked list wherein genes are characterized by their names as well as by a numerical score. We used two kinds of methods for re-ranking the genes selected by the enhanced RaJoLink rare-term algorithm described here: a) standard gene prioritization methods available via gene prioritization web servers (ToppGene and Endeavour [93,94] and b) propagation-based methods that were implemented on the STRINGS database [11], as briefly described below. In this formulation, a higher score indicates a better rank.

Re-ranking according to the personalized PageRank algorithm [33] is an iterative procedure where the score of genes at the (i+1)th iteration step is calculated as

$$p^{(i+1)} = (1-\alpha)M^T p^i + \alpha p^0 \quad (1)$$

where the α parameter is a constant in the range [0,1], and

$$m_{i,j} = \frac{a_{ij}}{\sum_{j=1}^{N} a_{ij}} \quad (2)$$

$\mathbf{p_i^0} = \frac{\mathbf{1}}{\mathbf{M}}$ $p_i^0 = 1/M$ if gene *i* belongs to the positive set, otherwise zero.

Re-ranking according to the diffusion kernel method [34] can be briefly written as follows:

$$p = e^{-tL_{\mu,\gamma}} p^0 \quad (3)$$

In this equation, the regularized Laplacian matrix is defined as

$$L_{\mu,\gamma} = QD - WAW \quad (4)$$

where the *W* and *Q* matrices are defined as follows:

$$w_{ij} = \begin{cases} \gamma \; if \; i = j \; and \; p_i^0 \neq 0 \\ 1 \; if \; i = j \; and \; p_i^0 = 0 \\ 0 \; i \neq j \end{cases} \quad (5)$$

Table 1. List of ovarian cancer biomarker genes used in this study.

	Symbol	Gene		Symbol	Gene		Symbol	Gene
1	CA125	CA 125 [37-40]	14	P16	p16 [41,42]	26	BIRC5	Survivin [43]
2	KRT19	Cytokeratin 19 [44,45]	15	CDKN1A	p21 [46-48]	27	TERT	hTERT [49]
3	KLK6	Kallikrein 6 [50]	16	CDKN1B	p27 [51-54]	28	EGFR	ERBB1 [55,56]
4	KLK10	Kallikrein 10 [57]	17	RB1	pRB [58,59]	29	ERBB2	ERBB2 [60]
5	IL6	Interleukin-6 [61]	18	E2F1	E2F1 [62]	30	MET	c-Met [63]
6	IL7	Interleukin-7 [64]	19	E2F2	E2F2 [65]	31	MMP2	MMP-2 [66]
7	IFNG	γ-interferon [67]	20	E2F4	E2F4 [65]	32	MMP9	MMP-9 [68]
8	FAS	sFas [69,70]	21	TP53	p53 [71,72]	33	MMP14	MT1-MMP [73]
9	VEGFR	VEGFR [74]	22	TP73	p73 [75]	34	WFDC2 (HE4)	Epididymis protein 4 [76-78]
10	CCND1	Cyclin D1 [46,79]	23	BAX	Bax [80,81]	35	SERPINB5	Maspin [82]
11	CCND3	Cyclin D3 [83]	24	BCL2L1	Bcl-xl [84]	36	BRCA1	BRCA1 [85]
12	CCNE	Cyclin E [86-89]	25	BIRC2	cIAP [90]	37	ERCC1	ERCC1 [91]
13	P15	p15 [92]						

$$q_{ij} = \begin{cases} \mu \; if \; i = j \; and \; p_i^0 \neq 0 \\ 1 \; if \; i = j \; and \; p_i^0 = 0 \\ 0 \; i \neq j \end{cases} \quad (6)$$

t,μ, and γ are the parameters of algorithm. Both algorithms were implemented in Matlab.

4. RESULTS

4.1. Including Gene Symbols into Hypothesis Generation: The Enhanced Rare-term Model

We investigated the issue of whether hypotheses about presently known gene - ovarian cancer associations could be generated by using a cut-off date. The open discovery process of the original rare-term model was guided with information about terms that are rarely connected to ovarian cancer and that represent genes according to the HUGO gene nomenclature database. We performed ovarian cancer experiments, both with the original and with the enhanced RaJoLink rare-term model. Also, abstracts of MEDLINE articles served as a source of biomedical information in both models, while the HUGO gene nomenclature knowledge base was used only in the enhanced RaJoLink rare-term model to guide the link analysis from the disease towards candidate genes, as shown in Fig. (**3**).

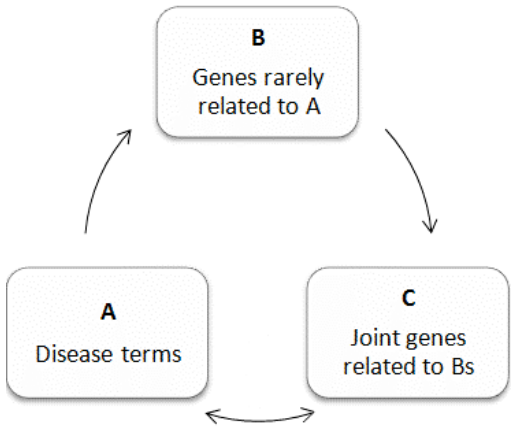

Figure 3. Hypothesis generation in the enhanced rare-term model.

The enhanced RaJoLink rare-term model is designed to finding candidates for the generation of hypotheses (i.e. the candidates for *C*) based on exploring gene names and symbols that rarely appear in the articles of the starting domain A (Fig. **3**). Compared to the original RaJoLink rare-term model [19] the enhanced model explores the open discovery setting (hypothesis generation) also from the point of view of disease-gene link discovery. Moreover, the new methodology is based on the assumption that by exploring terms related to an approved gene nomenclature, it should be faster to discover joint concepts that can reveal previously unknown links from literature *A* to literature *C*. In fact, the experimental results (Table **2**) show that the modified methodology improves the link discovery.

4.2. Combining Text Mining with Propagation Algorithms

Preliminary, exploratory experiments showed that the original RaJoLink rare-term model could not easily handle protein/gene names. We found that extending it with standardized protein/gene vocabulary leads to improvement; however we also noticed that some of the test genes did not rank well in the top of the list. We thus applied an additional ranking step in which the output of the text-mining step was re-ranked either with the Personalized PageRank algorithm (PDR), or the Personalized Diffusion Ranking algorithm (PR), or one of the standard gene-prioritization servers ToppGene or Endeavour [93,94]. As a result we obtained a number of candidate methods that we compared on the benchmark datasets. These were the original RaJoLink (OR), Enhanced RaJoLink (ER) text mining methods and the combined methods: ER + PR, ER + PDR, ER + Endeavour gene prioritization server [93,94] and ER + ToppGene gene prioritization server [93,94].

4.3 Testing the Methods on Ovarian Cancer Genes

Our test cases were based on a list of 37 genes associated with ovarian cancer (OC) (Table **1**) [12]. We designed two evaluation scenarios (4.3.1 and 4.3.2).

4.3.1. Testing the Methods on the Rediscovery of OC Biomarker Genes

We sought to establish whether the genes that have been proposed as OC biomarkers after 2007, could have been predicted on the basis of prior literature evidence and knowledge. We considered genes suggested as biomarkers as those that co-occurred with the term “marker” or “biomarker” in MEDLINE abstracts. We used MEDLINE abstracts, MeSH and HUGO terms published before 2007 (ovarian cancer test corpus, in Supplementary Datasets), and used standard propagation algorithms (PageRank or diffusion kernel methods [33,34]) for re-ranking the results, using the network of the STRING database, release version 6.3 (in use from December 12, 2005 to January 15, 2007). In the re-ranking step we could not use the gene-prioritization servers as the current servers contain information entered after 2007.

Out of the 37 ovarian cancer genes listed in Table **1**, 27 are mentioned together with “marker” or “biomarker” in MEDLINE articles published before 2007. The remaining 10 genes (our target genes) are: BCL2L1, CCND3, E2F1, E2F2, E2F4, ERCC1, IL7, MET, MMP9, WFDC2. Six genes were identified with the enhanced RajoLink method. For the five of these six genes, the ranks could be substantially improved by propagation/re-ranking (Table **2**). The full ranking is deposited as Supplementary Table 4.3.1.

4.3.2. Prediction of Ovarian Cancer Biomarkers Based on Current Knowledge

We wanted to establish if any putative gene biomarkers might exist for ovarian cancer on the basis of currently available published knowledge. To achieve this an experiment similar to the previous one was completed where i) the data input was the ovarian cancer prediction corpus (Supplementary Dataset 2) which includes abstracts about the known OC biomarker genes [12] (Table **1**) published up until May 14th

Table 2. Experiment 4.3.1: Rediscovery of genes suggested as OC biomarkers.

Gene symbol	Gene	Year when first mentioned as ovarian cancer prognostic marker	Rank1			
			Original RaJoLink	New RaJoLink	New RaJoLink + PageRank	New RaJoLink + Personal Diffusion
BCL2L1	Bcl-xl	2007 [84]	NA2	337	5	10
CCND3	Cyclin D3	2007 [83]	NA	165	43	31
E2F1	E2F1	2008 [62]	NA	NA	NA	NA
E2F2	E2F2	2007 [65]	69	140	36	3
E2F4	E2F4	2007 [65]	39	16	NA	NA
ERCC1	ERCC1	2007 [91]	NA	NA	NA	NA
IL7	Interleukin 7	2007 [64]	54	297	82	80
MET	c-Met	2007 [63]	NA	NA	NA	NA
MMP9	MMP-9	2007 [68]	44	86	22	49
WFDC2	Epididymis protein 4	2009 [76-78]	NA	NA	NA	NA

[1]A lower number indicates a better rank. [2]NA = not available

2012 and current versions of STRING, MeSH, HUGO nomenclature data, and ii) the propagation step was carried out with the standard propagation algorithms (PageRank or diffusion kernel methods [33,34]), and also with the gene prioritization servers ToppGene and Endeavour [93,94]. The ranks are presented in Supplementary table 4.3.2. It is apparent that a number of well/known cancer-related genes appear in the top of these lists.

For a better overview, we compared the top of the lists and picked 10 genes that ranked highly in most of the rankings (Table **3**). These include RUNX2, SOCS3, BCL6, PAX6, DAPK1, SMARCB1, RAF1, E2F6, P18INK4C (CDKN2C), and PAX5. These are all cancer-related genes that have not previously been proposed as OC biomarkers and have not been mentioned in literature sources together with ovarian cancer. These may represent genetic markers upon which hypotheses can be formulated in relation to ovarian cancer.

5. DISCUSSION

This work describes an open knowledge discovery methodology that generates hypotheses that are not known in advance. The methodology suggests potential disease-gene associations based on text databases (in our case MEDLINE), MeSH terms, and the HUGO nomenclature. In some cases, the methods presented here achieved suboptimal results, especially when abstracts contained many possible effectors. In these cases the simple, automated relation extraction gave erroneous results. Therefore we believe that the present approach can be further improved by integrating more sophisticated relation extraction methodologies. Above all, we suggest adapting the existing literature-based hypothesis generation techniques in semantic role labeling and biomedical relation extraction for the performance improvement. Semantic role labeling is a natural language processing technique for identifying the semantic roles of terms and phrases within sentences and to extract predicate-argument structures directly from texts. In combination with the biomedical relation mining these natural language processing approaches can be successfully exploited to recognize biomedical relations between different entities [95]. Nevertheless, we find that unequivocal matching to gene symbols is a very important factor, and that re-ranking text-based predictions either by standard propagation algorithms (PageRank, Diffusion Ranking) applied to the STRING network, or by gene-prioritization servers available on the web can improve the efficiency of text-mining searches.

Our searches revealed a number of genes that were previously not associated with progression and prognosis of ovarian cancer in MEDLINE abstracts. The *RUNX2* transcription factor is a putative tumor suppressor gene localized at chromosome *1p36*, a region showing frequent loss of heterozygosity events in colon, gastric, breast and ovarian cancers [96]. *RUNX2* has also been associated with prostate [97], lung [98], breast cancer [99], osteosarcoma [100] and thyroid tumors [101]. Several studies demonstrated a link between *RUNX2* and the hormonal system in prostate [102] and breast cancer [103]. All these data suggest a possible contribution of *RUNX2* to proliferation via enhancing the growth factor effects of sexual hormones. The potential of the gene in this association is supported by the prognostic power of hormone receptors in ovarian cancer [104].

BCL6 (B-cell CLL/lymphoma 6) is another transcription factor found to be frequently mutated in diffuse large-cell lymphoma. The gene was related not only to lymphomas [105] and leukemias [106] but also to progression to breast [107], gastric [108] and lung cancer [109]. Interestingly, both BCL6 [110] and RUNX2 [111] are influenced by prolactin secretion.

The tumor suppressor *DAPK1* (death-associated protein kinase 1) is one of the key regulators of the extrinsic apoptotic pathway [112]. Genetic variation of DAPK1 is associ-

Table 3. Predicted OC biomarker genes.

Genes predicted as ovarian cancer biomarkers	
Gene symbol	**Gene**
RUNX2	Runt-related transcription factor 2
SOCS3	Suppressor of cytokine signaling 3
BCL6	B-cell lymphoma 6 protein
PAX6	Paired box protein Pax-6
DAPK1	Death-associated protein kinase 1
SMARCB1	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily B member 1
RAF1	RAF proto-oncogene serine/threonine protein kinase
E2F6	Transcription factor E2F6
P18INK4C (CDKN2C)	Cyclin-dependent kinase 4 inhibitor C
PAX5	Paired box protein Pax-5

ated with survival in breast cancer [113]. The methylation status of DAPK1 contributed to stratifying colon cancer patients into subgroups with different prognosis [114]. The correlation between apopototic machinery and DAPK expression has just recently been validated in the ovarian cancer cell line OVCAR-3 [115]. Taken together, DAPK1 might be a potent prognostic marker for predicting apoptotic activity and also tumor progression in various cancer types including ovarian cancer.

According to literature review by MEDLINE, the remaining top-ten candidate genes, PAX6, E2F6, SMARCB1 and PAX5 also regulate gene transcription, while SOCS3, RAF1 and P18INK4C play a role in signal transduction pathways. In summary, the identified genes modulate key elements of molecular pathways of tumorigenesis and have already been associated with the progression in various malignancies.

While the identification of these genes provides validation for the proposed methodology, we must emphasize the overall goal of our study: to provide a framework for future text mining approaches. Thus, the suggested pathway can be used for other diseases and for other databases as well.

6. CONCLUSIONS

Capturing information from heterogeneous data is one of the key problems of literature-based hypothesis generation. The strategy presented here is based on guiding the open knowledge discovery process by a standardized gene/protein nomenclature as well as by improving the ranking using gene-prioritization techniques. The results suggest that this combined strategy can enhance the efficiency of state-of-the-art methods for the literature-based discovery. We found that this strategy makes biologically meaningful candidate hypotheses i.e. the genes/proteins suggested as potential ovarian cancer biomarkers are known to be associated with various cancer-related mechanisms. Among them, the RUNX2 transcription factor, the DAPK1 tumor suppressor and the BCL6 transcription factor were identified as the most frequent genes extracted from literature abstracts related to the ovarian cancer biomarker genes.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

I. Petrič, B. Ligeti, B. Győrffy and S. Pongor designed the project together. I. Petrič designed and programmed the extended RaJoLink algorithm, B.Ligeti designed and programmed the re-ranking algorithms. I. Petrič and B. Ligeti carried out the calculations. I. Petrič, B. Ligeti, B. Győrffy and S. Pongor evaluated the results and wrote the manuscript. I.Petrič acknowledges the financial support of CEI Fellowship Programme CERES. Work at *Pázmány Péter Catholic University, Budapest* was partially supported by grants TÁMOP-4.2.1.B-11/2/KMR-2011-0002, TÁMOP-4.2.2/B-10/1-2010-0014 and TÉT 10-1-2011-0058. B. Győrffy was supported by the OTKA PD 83154 grant.

REFERENCES

[1] Cohen, A.M.; Hersh, W.R. A survey of current work in biomedical text mining. *Brief. Bioinform*, **2005**, *6*(1), 57-71.

[2] Srinivasan, P. Text Mining: Generating Hypotheses from MEDLINE. *J. Am. Soc. Inf. Sci. Tech*., **2004**, *55*(5), 396-413.

[3] Senator, T. Evidence extraction and link discovery program. Speech at DARPATech 2002. 2002.

[4] Agarwal, P.; Searls, D. Literature mining in support of drug discovery. *Brief. Bioinform*., **2008**, *9*(6), 479-492.

[5] Faro, A.; Giordano, D.; Spampinato, C. Combining literature text mining with microarray data: advances for system biology modeling. *Brief. Bioinform*., **2012**, *3*(1), 61-82.

[6] Barbosa-Silva, A.; Soldatos, T.; Magalhães, I.; Pavlopoulos, G.; Fontaine, J.; Andrade-Navarro, M.; Schneider, R.; JM, O. LAITOR - Literature Assistant for Identification of Terms co-Occurrences and Relationships. *BMC Bioinformatics*, **2010**, *11*, 10.

[7] Bell, L.; Chowdhary, R.; Liu, J.; Niu, X.; Zhang, J. Integrated Bio-Entity Network: A System for Biological Knowledge Discovery. *Plos One*, **2011**, *6*(6), e21474.

[8] Van Landeghem, S.; De Baets, B.; Van de Peer, Y.; Saeys, Y. High-precision bio-molecular event extraction from text using parallel binary classifiers. *Comput. Intel*., **2011**, *27*(4), 645-664.

[9] Petrič, I.; Urbančič, T.; Cestnik, B. Discovering hidden knowledge from biomedical literature. *Informatica*, **2007**, *31*(1), 15-20.

[10] Seal, R.L.; Gordon, S.M.; Lush, M.J.; Wright, M.W.; Bruford, E.A. genenames.org: the HGNC resources in 2011. *Nucleic Acids Res*., **2011**, *39*(suppl 1), D514-D519.

[11] Szklarczyk, D.; Franceschini, A.; Kuhn, M.; Simonovic, M.; Roth, A.; Minguez, P.; Doerks, T.; Stark, M.; Muller, J.; Bork, P.; Jensen, L.J.; Mering, C.V. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*., **2011**, *39*(Database issue), 561-568.

[12] Gyorffy, B.; Lánczky, A.; Szállási, Z. Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients. *Endocr. Relat. Cancer*, **2012**, *19*(2), 197-208.

[13] Swanson, D.R. Undiscovered public knowledge. *Libr. Q.*, **1986**, *56*(2), 103-118.

[14] Smalheiser, N.R.; Swanson, D. R. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput. Methods Programs Biomed*., **1998**, *57*(3), 149-153.

[15] Weeber, M.; Vos, R.; Klein, H.; Berg, L.T.W.d.J.-v.d. Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. *J. Am. Soc. Inf. Sci. Tech*., **2001**, *52*(7), 548-557.

[16] Swanson, D.R. Medical literature as a potential source of new knowledge. *Bull. Med. Libr. Assoc*., **1990**, *78*(1), 29-37.

[17] Lindsay, R.K.; Gordon, M.D. Literature-based discovery by lexical statistics. *J. Am. Soc. Inf. Sci. Tech*., **1999**, *50*(7), 574-587.

[18] Blake, C.; Pratt, W. In *Automatically Identifying Candidate Treatments from Existing Medical Literature*, AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases, Stanford, CA, Harabagiu, S.; Vinay, C., Eds. Stanford, CA, **2002**.

[19] Petrič, I.; Urbančič, T.; Cestnik, B.; Macedoni-Lukšič, M. Literature mining method RaJoLink for uncovering relations between biomedical concepts. *J. Biomed. Inform*., **2009**, *42*(2), 219-227.

[20] Urbančič, T.; Petrič, I.; Cestnik, B. In *RaJoLink: a method for finding seeds of future discoveries in nowadays literature*, Foundations of Intelligent Systems: 18th International Symposium, ISMIS 2009, Prague, Czech Republic, September 14-17, 2009; Rauch, J., Ed. Springer, Berlin; Heidelberg: Prague, Czech Republic, **2009**; pp 129-138.

[21] Nelson, S.J.; Johnston, D.; Humphreys, B.L.; Relationships in Medical Subject Headings. In *Relationships in the organization of knowledge*, Bean, C.A.; Green, R., Eds. Kluwer Academic Publishers: New York, **2001**; pp 171-184.

[22] Eyre, T.A.; Ducluzeau, F.; Sneddon, T.P.; Povey, S.; Bruford, E.A.; Lush, M.J. The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res*., **2006**, *34*(Database issue), D319-21.

[23] Szklarczyk, D.; Franceschini, A.; Kuhn, M.; Simonovic, M.; Roth, A.; Minguez, P.; Doerks, T.; Stark, M.; Muller, J.; Bork, P.; Jensen, L.J.; von Mering, C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*., **2011**, *39*((Database issue)), D561-D568.

[24] von Mering, C.; Huynen, M.; Jaeggi, D.; Schmidt, S.; Bork, P.; Snel, B. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*., **2003**, *31*(1), 258-261.

[25] Krallinger, M.; Valencia, A.; Hirschman, L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol*., **2008**, *9*(Suppl 2), S8.

[26] Grimes, G.; Wen, T.; Mewissen, M.; Baxter, R.; Moodie, S.; Beattie, J.; Ghazal, P. PDQ Wizard: automated prioritization and characterization of gene and protein lists using biomedical literature. *Bioinformatics*, **2006**, *22*(16), 2055-2057.

[27] Erhardt, R.A.-A.; Schneider, R.; Blaschke, C. Status of text-mining techniques applied to biomedical text. *Drug Discov. Today*, **2006**, *11*(7/8), 315-325.

[28] Jensen, L.; Saric, J.; Bork, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet*., **2006**, *7*, 119-129.

[29] Hristovski, D.; Kastrin, A.; Peterlin, B.; Rindflesch, T., Combining semantic relations and DNA microarray data for novel hypotheses generation. In *Linking Literature, Information, and Knowledge for Biology*, Blaschke, C.; Shatkay, H., Eds. Springer: **2010**; Vol. 6004, pp 53-61.

[30] Frijters, R.; van Vugt, M.; Smeets, R.; van Schaik, R.; de Vlieg, J.; Alkema, W. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput. Biol*., **2010**, *6*(9), e1000943.

[31] Hardy, J.; Singleton, A. Genomewide association studies and human disease. *N Engl. J. Med*., **2009**, *360*(17), 1759-1768.

[32] Tranchevent, L.C.; Capdevila, F.B.; Nitsch, D.; De Moor, B.; De Causmaecker, P.Y.M. A guide to web tools to prioritize candidate genes. *Brief Bioinform*., **2011**, *12*(1), 22-32.

[33] Brin, S.; Page, L. The anatomy of a large-scale hypertextual web search engine. *Comp. Net. ISDN Sys*., **1998**, *30*(1-7), 107-117.

[34] Kondor, R.I.; Lafferty, J. In *Diffusion kernels on graphs and other discrete input spaces*, ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning, San Francisco, CA, USA, Morgan Kaufmann: San Francisco, CA, USA, **2002**; pp 315-322.

[35] U.S. National Library of Medicine PubMed Overview. http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/index.html.

[36] Sayers, E.; Wheeler, D., Building Customized Data Pipelines Using the Entrez Programming Utilities (eUtils). In: *NCBI Short Courses* [Online] U.S. National Center for Biotechnology Information 2004. http://www.ncbi.nlm.nih.gov/books/NBK1056/.

[37] Cooper, B.C.; Sood, A.K.; Davis, C.S.; Ritchie, J.M.; Sorosky, J.I.; Anderson, B.; Buller, R.E. Preoperative CA 125 levels: an independent prognostic factor for epithelial ovarian cancer. *Obstet. Gynecol*., **2002**, *100*(1), 59-64.

[38] Gadducci, A.; Cosio, S.; Fanucchi, A.; Negri, S.; Cristofani, R.; Genazzani, A.R. The predictive and prognostic value of serum CA 125 half-life during paclitaxel/platinum-based chemotherapy in patients with advanced ovarian carcinoma. *Gynecol. Oncol*., **2004**, *93*(1), 131-6.

[39] Gadducci, A.; Zola, P.; Landoni, F.; Maggino, T.; Sartori, E.; Bergamino, T.; Cristofani, R. Serum half-life of CA 125 during early chemotherapy as an independent prognostic variable for patients with advanced epithelial ovarian cancer: results of a multicentric Italian study. *Gynecol. Oncol*., **1995**, *58*(1), 42-7.

[40] Riedinger, J.M.; Wafflart, J.; Ricolleau, G.; Eche, N.; Larbre, H.; Basuyau, J.P.; Dalifard, I.; Hacene, K.; Pichon, M.F. CA 125 half-life and CA 125 nadir during induction chemotherapy are independent predictors of epithelial ovarian cancer outcome: results of a French multicentric study. *Ann. Oncol*., **2006**, *17*(8), 1234-8.

[41] Katsaros, D.; Cho, W.; Singal, R.; Fracchioli, S.; Rigault De La Longrais, I.A.; Arisio, R.; Massobrio, M.; Smith, M.; Zheng, W.; Glass, J.; Yu, H. Methylation of tumor suppressor gene p16 and prognosis of epithelial ovarian cancer. *Gynecol. Oncol*., **2004**, *94*(3), 685-92.

[42] Kommoss, S.; du Bois, A.; Ridder, R.; Trunk, M.J.; Schmidt, D.; Pfisterer, J.; Kommoss, F. Independent prognostic significance of cell cycle regulator proteins p16(INK4a) and pRb in advanced-stage ovarian carcinoma including optimally debulked patients: a translational research subprotocol of a randomised study of the Arbeitsgemeinschaft Gynaekologische Onkologie Ovarian Cancer Study Group. *Br. J. Cancer*, **2007**, *96*(2), 306-13.

[43] Sui, L.; Dong, Y.; Ohno, M.; Watanabe, Y.; Sugimoto, K.; Tokuda, M. Survivin expression and its correlation with cell proliferation and prognosis in epithelial ovarian tumors. *Int. J. Oncol*., **2002**, *21*(2), 315-20.

[44] Gadducci, A.; Ferdeghini, M.; Cosio, S.; Fanucchi, A.; Cristofani, R.; Genazzani, A.R. The clinical relevance of serum CYFRA 21-1 assay in patients with ovarian cancer. *Int. J. Gynecol. Cancer*, **2001**, *11*(4), 277-82.

[45] Tempfer, C.; Hefler, L.; Heinzl, H.; Loesch, A.; Gitsch, G.; Rumpold, H.; Kainz, C. CYFRA 21-1 serum levels in women with adnexal masses and inflammatory diseases. *Br. J. Cancer*, **1998**, *78*(8), 1108-12.

[46] Bali, A.; O'Brien, P.M.; Edwards, L.S.; Sutherland, R.L.; Hacker, N.F.; Henshall, S.M. Cyclin D1, p53, and p21Waf1/Cip1 expression is predictive of poor clinical outcome in serous epithelial ovarian cancer. *Clin. Cancer Res*., **2004**, *10*(15), 5168-77.

[47] Ferrandina, G.; Stoler, A.; Fagotti, A.; Fanfani, F.; Sacco, R.; De Pasqua, A.; Mancuso, S.; Scambia, G. p21WAF1/CIP1 protein expression in primary ovarian cancer. *Int. J. Oncol*., **2000**, *17*(6), 1231-5.

[48] Plisiecka-Halasa, J.; Karpinska, G.; Szymanska, T.; Ziolkowska, I.; Madry, R.; Timorek, A.; Debniak, J.; Ulanska, M.; Jedryka, M.; Chudecka-Glaz, A.; Klimek, M.; Rembiszewska, A.; Kraszewska, E.; Dybowski, B.; Markowska, J.; Emerich, J.; Pluzanska, A.; Goluda, M.; Rzepka-Gorska, I.; Urbanski, K.; Zielinski, J.; Stelmachow, J.; Chrabowska, M.; Kupryjanczyk, J. P21WAF1, P27KIP1, TP53 and C-MYC analysis in 204 ovarian carcinomas treated with platinum-based regimens. *Ann. Oncol.*, **2003**, *14*(7), 1078-85.

[49] Brustmann, H. Immunohistochemical detection of human telomerase reverse transcriptase (hTERT) and c-kit in serous ovarian carcinoma: a clinicopathologic study. *Gynecol. Oncol.*, **2005**, *98*(3), 396-402.

[50] Diamandis, E.P.; Scorilas, A.; Fracchioli, S.; Van Gramberen, M.; De Bruijn, H.; Henrik, A.; Soosaipillai, A.; Grass, L.; Yousef, G.M.; Stenman, U.H.; Massobrio, M.; Van Der Zee, A.G.; Vergote, I.; Katsaros, D. Human kallikrein 6 (hK6): a new potential serum biomarker for diagnosis and prognosis of ovarian carcinoma. *J. Clin. Oncol.*, **2003**, *21*(6), 1035-43.

[51] Korkolopoulou, P.; Vassilopoulos, I.; Konstantinidou, A.E.; Zorzos, H.; Patsouris, E.; Agapitos, E.; Davaris, P. The combined evaluation of p27Kip1 and Ki-67 expression provides independent information on overall survival of ovarian carcinoma patients. *Gynecol. Oncol.*, **2002**, *85*(3), 404-14.

[52] Masciullo, V.; Ferrandina, G.; Pucci, B.; Fanfani, F.; Lovergine, S.; Palazzo, J.; Zannoni, G.; Mancuso, S.; Scambia, G.; Giordano, A. p27Kip1 expression is associated with clinical outcome in advanced epithelial ovarian cancer: multivariate analysis. *Clin. Cancer Res.*, **2000**, *6*(12), 4816-22.

[53] Newcomb, E.W.; Sosnow, M.; Demopoulos, R.I.; Zeleniuch-Jacquotte, A.; Sorich, J.; Speyer, J.L. Expression of the cell cycle inhibitor p27KIP1 is a new prognostic marker associated with survival in epithelial ovarian tumors. *Am. J. Pathol.*, **1999**, *154*(1), 119-25.

[54] Schmider-Ross, A.; Pirsig, O.; Gottschalk, E.; Denkert, C.; Lichtenegger, W.; Reles, A. Cyclin-dependent kinase inhibitors CIP1 (p21) and KIP1 (p27) in ovarian cancer. *J. Cancer Res. Clin. Oncol.*, **2006**, *132*(3), 163-70.

[55] Skirnisdottir, I.; Seidal, T.; Sorbe, B. A new prognostic model comprising p53, EGFR, and tumor grade in early stage epithelial ovarian carcinoma and avoiding the problem of inaccurate surgical staging. *Int. J. Gynecol. Cancer*, **2004**, *14*(2), 259-70.

[56] Psyrri, A.; Kassar, M.; Yu, Z.; Bamias, A.; Weinberger, P.M.; Markakis, S.; Kowalski, D.; Camp, R.L.; Rimm, D.L.; Dimopoulos, M.A. Effect of epidermal growth factor receptor expression level on survival in patients with epithelial ovarian cancer. *Clin. Cancer Res.*, **2005**, *11*(24 Pt 1), 8637-43.

[57] Luo, L.Y.; Katsaros, D.; Scorilas, A.; Fracchioli, S.; Piccinno, R.; Rigault de la Longrais, I.A.; Howarth, D.J.; Diamandis, E.P. Prognostic value of human kallikrein 10 expression in epithelial ovarian carcinoma. *Clin. Cancer Res.*, **2001**, *7*(8), 2372-9.

[58] Dong, Y.; Walsh, M.D.; McGuckin, M.A.; Cummings, M.C.; Gabrielli, B.G.; Wright, G.R.; Hurst, T.; Khoo, S.K.; Parsons, P.G. Reduced expression of retinoblastoma gene product (pRB) and high expression of p53 are associated with poor prognosis in ovarian cancer. *Int. J. Cancer*, **1997**, *74*(4), 407-15.

[59] Konstantinidou, A.E.; Korkolopoulou, P.; Vassilopoulos, I.; Tsenga, A.; Thymara, I.; Agapitos, E.; Patsouris, E.; Davaris, P. Reduced retinoblastoma gene protein to Ki-67 ratio is an adverse prognostic indicator for ovarian adenocarcinoma patients. *Gynecol. Oncol.*, **2003**, *88*(3), 369-78.

[60] Lassus, H.; Leminen, A.; Vayrynen, A.; Cheng, G.; Gustafsson, J.A.; Isola, J.; Butzow, R. ERBB2 amplification is superior to protein expression status in predicting patient outcome in serous ovarian carcinoma. *Gynecol. Oncol.*, **2004**, *92*(1), 31-9.

[61] Scambia, G.; Testa, U.; Benedetti Panici, P.; Foti, E.; Martucci, R.; Gadducci, A.; Perillo, A.; Facchini, V.; Peschle, C.; Mancuso, S. Prognostic significance of interleukin 6 serum levels in patients with ovarian cancer. *Br. J. Cancer*, **1995**, *71*(2), 354-6.

[62] Suh, D.S.; Yoon, M.S.; Choi, K.U.; Kim, J.Y. Significance of E2F-1 overexpression in epithelial ovarian cancer. *Int. J. Gynecol. Cancer*, **2008**, *18*(3), 492-8.

[63] Sawada, K.; Radjabi, A.R.; Shinomiya, N.; Kistner, E.; Kenny, H.; Becker, A.R.; Turkyilmaz, M.A.; Salgia, R.; Yamada, S.D.; Vande Woude, G.F.; Tretiakova, M.S.; Lengyel, E. c-Met overexpression is a prognostic factor in ovarian cancer and an effective target for inhibition of peritoneal dissemination and invasion. *Cancer Res.*, **2007**, *67*(4), 1670-9.

[64] Lambeck, A.J.; Crijns, A.P.; Leffers, N.; Sluiter, W.J.; ten Hoor, K.A.; Braid, M.; van der Zee, A.G.; Daemen, T.; Nijman, H.W.; Kast, W.M. Serum cytokine profiling as a diagnostic and prognostic tool in ovarian cancer: a potential role for interleukin 7. *Clin. Cancer Res.*, **2007**, *13*(8), 2385-91.

[65] Reimer, D.; Sadr, S.; Wiedemair, A.; Stadlmann, S.; Concin, N.; Hofstetter, G.; Muller-Holzner, E.; Marth, C.; Zeimet, A.G. Clinical relevance of E2F family members in ovarian cancer--an evaluation in a training set of 77 patients. *Clin. Cancer Res.*, **2007**, *13*(1), 144-51.

[66] Torng, P.L.; Mao, T.L.; Chan, W.Y.; Huang, S.C.; Lin, C.T. Prognostic significance of stromal metalloproteinase-2 in ovarian adenocarcinoma and its relation to carcinoma progression. *Gynecol. Oncol.*, **2004**, *92*(2), 559-67.

[67] Marth, C.; Fiegl, H.; Zeimet, A.G.; Muller-Holzner, E.; Deibl, M.; Doppler, W.; Daxenbichler, G. Interferon-gamma expression is an independent prognostic factor in ovarian cancer. *Am. J. Obstet. Gynecol.*, **2004**, *191*(5), 1598-605.

[68] Sillanpaa, S.; Anttila, M.; Voutilainen, K.; Ropponen, K.; Turpeenniemi-Hujanen, T.; Puistola, U.; Tammi, R.; Tammi, M.; Sironen, R.; Saarikoski, S.; Kosma, V.M. Prognostic significance of matrix metalloproteinase-9 (MMP-9) in epithelial ovarian cancer. *Gynecol. Oncol.*, **2007**, *104*(2), 296-303.

[69] Hefler, L.; Mayerhofer, K.; Nardi, A.; Reinthaller, A.; Kainz, C.; Tempfer, C. Serum soluble Fas levels in ovarian cancer. *Obstet. Gynecol.*, **2000**, *96*(1), 65-9.

[70] Konno, R.; Takano, T.; Sato, S.; Yajima, A. Serum soluble fas level as a prognostic factor in patients with gynecological malignancies. *Clin. Cancer Res.*, **2000**, *6*(9), 3576-80.

[71] Buttitta, F.; Marchetti, A.; Gadducci, A.; Pellegrini, S.; Morganti, M.; Carnicelli, V.; Cosio, S.; Gagetti, O.; Genazzani, A. R.; Bevilacqua, G. p53 alterations are predictive of chemoresistance and aggressiveness in ovarian carcinomas: a molecular and immunohistochemical study. *Br. J. Cancer*, **1997**, *75*(2), 230-5.

[72] Reles, A.; Wen, W.H.; Schmider, A.; Gee, C.; Runnebaum, I.B.; Kilian, U.; Jones, L.A.; El-Naggar, A.; Minguillon, C.; Schonborn, I.; Reich, O.; Kreienberg, R.; Lichtenegger, W.; Press, M.F. Correlation of p53 mutations with resistance to platinum-based chemotherapy and shortened survival in ovarian cancer. *Clin. Cancer Res.*, **2001**, *7*(10), 2984-97.

[73] Kamat, A.A.; Fletcher, M.; Gruman, L.M.; Mueller, P.; Lopez, A.; Landen, C.N. Jr.; Han, L.; Gershenson, D.M.; Sood, A.K. The clinical relevance of stromal matrix metalloproteinase expression in ovarian cancer. *Clin. Cancer Res.*, **2006**, *12*(6), 1707-14.

[74] Hefler, L.A.; Zeillinger, R.; Grimm, C.; Sood, A.K.; Cheng, W.F.; Gadducci, A.; Tempfer, C.B.; Reinthaller, A. Preoperative serum vascular endothelial growth factor as a prognostic parameter in ovarian cancer. *Gynecol. Oncol.*, **2006**, *103*(2), 512-7.

[75] Becker, K.; Pancoska, P.; Concin, N.; Vanden Heuvel, K.; Slade, N.; Fischer, M.; Chalas, E.; Moll, U.M. Patterns of p73 N-terminal isoform expression and p53 status have prognostic value in gynecological cancers. *Int. J. Oncol.*, **2006**, *29*(4), 889-902.

[76] Huhtinen, K.; Suvitie, P.; Hiissa, J.; Junnila, J.; Huvila, J.; Kujari, H.; Setala, M.; Harkki, P.; Jalkanen, J.; Fraser, J.; Makinen, J.; Auranen, A.; Poutanen, M.; Perheentupa, A. Serum HE4 concentration differentiates malignant ovarian tumours from ovarian endometriotic cysts. *Br. J. Cancer*, **2009**, *100*(8), 1315-9.

[77] Moore, R.G.; Jabre-Raughley, M.; Brown, A.K.; Robison, K.M.; Miller, M.C.; Allard, W.J.; Kurman, R.J.; Bast, R.C.; Skates, S.J. Comparison of a novel multiple marker assay vs the Risk of Malignancy Index for the prediction of epithelial ovarian cancer in patients with a pelvic mass. *Am. J. Obstet. Gynecol.*, **2010**, *203*(3), 228 e1-6.

[78] Moore, R.G.; McMeekin, D.S.; Brown, A.K.; DiSilvestro, P.; Miller, M.C.; Allard, W.J.; Gajewski, W.; Kurman, R.; Bast, R.C., Jr.; Skates, S.J. A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass. *Gynecol. Oncol.*, **2009**, *112*(1), 40-6.

[79] Barbieri, F.; Lorenzi, P.; Ragni, N.; Schettini, G.; Bruzzo, C.; Pedulla, F.; Alama, A. Overexpression of cyclin D1 is associated with poor survival in epithelial ovarian cancer. *Oncology*, **2004**, *66*(4), 310-5.

[80] Skirnisdottir, I.; Sorbe, B.; Seidal, T. P53, bcl-2, and bax: their relationship and effect on prognosis in early stage epithelial ovarian carcinoma. *Int. J. Gynecol. Cancer*, **2001**, *11*(2), 147-58.

[81] Tai, Y.T.; Lee, S.; Niloff, E.; Weisman, C.; Strobel, T.; Cannistra, S.A. BAX protein expression and clinical outcome in epithelial ovarian cancer. *J. Clin. Oncol.*, **1998**, *16*(8), 2583-90.

[82] Secord, A.A.; Lee, P.S.; Darcy, K.M.; Havrilesky, L.J.; Grace, L.A.; Marks, J.R.; Berchuck, A. Maspin expression in epithelial ovarian cancer and associations with poor prognosis: a Gynecologic Oncology Group study. *Gynecol. Oncol.*, **2006**, *101*(3), 390-7.

[83] Levidou, G.; Korkolopoulou, P.; Thymara, I.; Vassilopoulos, I.; Saetta, A.A.; Gakiopoulou, H.; Konstantinidou, A.; Kairi-Vassilatou, E.; Pavlakis, K.; Patsouris, E. Expression and prognostic significance of cyclin D3 in ovarian adenocarcinomas. *Int. J. Gynecol. Pathol.*, **2007**, *26*(4), 410-7.

[84] Materna, V.; Surowiak, P.; Markwitz, E.; Spaczynski, M.; Drag-Zalesinska, M.; Zabel, M.; Lage, H. Expression of factors involved in regulation of DNA mismatch repair- and apoptosis pathways in ovarian cancer patients. *Oncol. Rep.*, **2007**, *17*(3), 505-16.

[85] Thrall, M.; Gallion, H.H.; Kryscio, R.; Kapali, M.; Armstrong, D.K.; DeLoia, J.A. BRCA1 expression in a large series of sporadic ovarian carcinomas: a Gynecologic Oncology Group study. *Int. J. Gynecol. Cancer*, **2006**, *16*(Suppl 1), 166-71.

[86] Bedrosian, I.; Lee, C.; Tucker, S.L.; Palla, S.L.; Lu, K.; Keyomarsi, K. Cyclin E-associated kinase activity predicts response to platinum-based chemotherapy. *Clin. Cancer Res.*, **2007**, *13*(16), 4800-6.

[87] Farley, J.; Smith, L.M.; Darcy, K.M.; Sobel, E.; O'Connor, D.; Henderson, B.; Morrison, L.E.; Birrer, M.J. Cyclin E expression is a significant predictor of survival in advanced, suboptimally debulked ovarian epithelial cancers: a Gynecologic Oncology Group study. *Cancer Res.*, **2003**, *63*(6), 1235-41.

[88] Rosen, D. G.; Yang, G.; Deavers, M. T.; Malpica, A.; Kavanagh, J. J.; Mills, G. B.; Liu, J. Cyclin E expression is correlated with tumor progression and predicts a poor prognosis in patients with ovarian carcinoma. *Cancer* **2006,** *106* (9), 1925-32.

[89] Sui, L.; Dong, Y.; Ohno, M.; Sugimoto, K.; Tai, Y.; Hando, T.; Tokuda, M. Implication of malignancy and prognosis of p27(kip1), Cyclin E, and Cdk2 expression in epithelial ovarian tumors. *Gynecol. Oncol.*, **2001**, *83*(1), 56-63.

[90] Psyrri, A.; Yu, Z.; Bamias, A.; Weinberger, P.M.; Markakis, S.; Kowalski, D.; Camp, R.L.; Rimm, D. L.; Dimopoulos, M.A. Evaluation of the prognostic value of cellular inhibitor of apoptosis protein in epithelial ovarian cancer using automated quantitative protein analysis. *Cancer Epidemiol. Biomarkers Prev.*, **2006**, *15*(6), 1179-83.

[91] Darcy, K.M.; Tian, C.; Reed, E. A Gynecologic Oncology Group study of platinum-DNA adducts and excision repair cross-complementation group 1 expression in optimal, stage III epithelial ovarian cancer treated with platinum-taxane chemotherapy. *Cancer Res.*, **2007**, *67*(9), 4474-81.

[92] Kudoh, K.; Ichikawa, Y.; Yoshida, S.; Hirai, M.; Kikuchi, Y.; Nagata, I.; Miwa, M.; Uchida, K. Inactivation of p16/CDKN2 and p15/MTS2 is associated with prognosis and response to chemotherapy in ovarian cancer. *Int. J. Cancer*, **2002**, *99*(4), 579-82.

[93] Chen, J.; Bardes, E.E.; Aronow, B.J.; Jegga, A.G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.*, **2009**, *37*(Web Server issue), W305-11.

[94] Tranchevent, L.C.; Barriot, R.; Yu, S.; Van Vooren, S.; Van Loo, P.; Coessens, B.; De Moor, B.; Aerts, S.; Moreau, Y. ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res.*, **2008**, *36*(Web Server issue), W377-84.

[95] Tymoshenko, K.; Somasundaran, S.; Prabhakaran, V.; Shet, V. Relation Mining in the Biomedical Domain using Entity-level Semantics. *Front Art Intel Apllic*, **2012**, *242*, 780-785.

[96] Goel, A.; Arnold, C.; Tassone, P.; Chang, D.; Niedzwiecki, D.; Dowell, J.; Wasserman, L.; Compton, C.; Mayer, R.; Bertagnolli, M.M.; Boland, C. Epigenetic inactivation of RUNX3 in microsatellite unstable sporadic colon cancers. *Int. J. Cancer*, **2004**, *112*(5), 754-759.

[97] Little, G.; Noushmehr, H.; Baniwal, S.; Berman, B.; Coetzee, G.; Frenkel, B. Genome-wide Runx2 occupancy in prostate cancer cells suggests a role in regulating secretion. *Nucleic Acids Res.*, **2012**, *40*(8), 3538-3547.

[98] Tandon, M.; Gokul, K.; Ali, S.; Chen, Z.; Lian, J.; Stein, G.; Pratap, J. Runx2 mediates epigenetic silencing of the bone morphogenetic protein-3B (BMP-3B/GDF10) in lung cancer cells. *Mol. Cancer*, **2012**, *11*, 27.

[99] Chimge, N.; Baniwal, S.; Little, G.; Chen, Y.; Kahn, M.; Tripathy, D.; Borok, Z.; Frenkel, B. Regulation of breast cancer metastasis by Runx2 and estrogen signaling: the role of SNAI2. *Breast Cancer Res* **2011**, *13*(6), R127.

[100] Martin, J.; Zielenska, M.; Stein, G.; van Wijnen, A.; Squire, J. The Role of RUNX2 in Osteosarcoma Oncogenesis. *Sarcoma*, **2011**, *282745*.

[101] Dalle Carbonare, L.; Frigo, A.; Francia, G.; Davì, M.; Donatelli, L.; Stranieri, C.; Brazzarola, P.; Zatelli, M.; Menestrina, F.; Valenti, M. Runx2 mRNA Expression in the Tissue, Serum, and Circulating Non-Hematopoietic Cells of Patients with Thyroid Cancer. *J. Clin. Endocrinol. Metab.*, **2012**, *97*(7), E1249-56.

[102] van der Deen, M.; Akech, J.; Wang, T.; FitzGerald, T.; Altieri, D.; Languino, L.; Lian, J.; van Wijnen, A.; Stein, J.; Stein, G. The cancer-related Runx2 protein enhances cell growth and responses to androgen and TGFbeta in prostate cancer cells. *J. Cell Biochem.*, **2010**, *109*(4), 828-837.

[103] Pratap, J.; Wixted, J.; Gaur, T.; Zaidi, S.; Dobson, J.; Gokul, K.; Hussain, S.; van Wijnen, A.; Stein, J.; Stein, G.; Lian, J. Runx2 transcriptional activation of Indian Hedgehog and a downstream bone metastatic pathway in breast cancer cells. *Cancer Res.*, **2008**, *68*(19), 7795-7802.

[104] Fekete, T.; Rásó, E.; Pete, I.; Tegze, B.; Liko, I.; Munkácsy, G.; Sipos, N.; Rigó, J.J.; Györffy, B. Meta-analysis of gene expression profiles associated with histological classification and survival in 829 ovarian cancer samples. *Int. J. Cancer*, **2012**, *131*(1), 95-105.

[105] Wagner, S.; Ahearne, M.; Ko Ferrigno, P. The role of BCL6 in lymphomas and routes to therapy. *Br. J. Haematol.*, **2011**, *152*(1), 3-12.

[106] Pellicano, F.; Holyoake, T. Assembling defenses against therapy-resistant leukemic stem cells: Bcl6 joins the ranks. *J. Exp. Med.*, **2011**, *208*(11), 2155-2158.

[107] Pinto, A.; André, S.; Silva, G.; Vieira, S.; Santos, A.; Dias, S.; Soares, J. BCL-6 oncoprotein in breast cancer: loss of expression in disease progression. *Pathobiology*, **2009**, *76*(5), 235-242.

[108] Hirata, Y.; Ogasawara, N.; Sasaki, M.; Mizushima, T.; Shimura, T.; Mizoshita, T.; Mori, Y.; Kubota, E.; Wada, T.; Tanida, S.; Kataoka, H.; Kamiya, T.; Higashiyama, S.; Joh, T. BCL6 degradation caused by the interaction with the C-terminus of pro-HB-EGF induces cyclin D2 expression in gastric cancers. *Br. J. Cancer*, **2009**, *100*(8), 1320-1329.

[109] Dmitriev, A.; Kashuba, V.; Haraldson, K.; Senchenko, V.; Pavlova, T.; Kudryavtseva, A.; Anedchenko, E.; Krasnov, G.; Pronina, I.; Loginov, V.; Kondratieva, T.; Kazubskaya, T.; Braga, E.; Yenamandra, S.; Ignatjev, I.; Ernberg, I.; Klein, G.; Lerman, M.; Zabarovsky, E. Genetic and epigenetic analysis of non-small cell lung cancer with NotI-microarrays. *Epigenetics*, **2012**, *7*(5), 502-513.

[110] Tran, T.; Utama, F.; Lin, J.; Yang, N.; Sjolund, A.; Ryder, A.; Johnson, K.; Neilson, L.; Liu, C.; Brill, K.; Rosenberg, A.; Witkiewicz, A.; Rui, H. Prolactin inhibits BCL6 expression in breast cancer through a Stat5a-dependent mechanism. *Cancer Res.*, **2010**, *70*(4), 1711-1721.

[111] Charoenphandhu, N.; Teerapornpuntakit, J.; Methawasin, M.; Wongdee, K.; Thongchote, K.; Krishnamra, N. Prolactin decreases expression of Runx2, osteoprotegerin, and RANKL in primary osteoblasts derived from tibiae of adult female rats. *Can. J. Physiol. Pharmacol.*, **2008**, *86*(5), 240-248.

[112] Aberg, M.; Johnell, M.; Wickström, M.; Siegbahn, A. Tissue Factor/ FVIIa prevents the extrinsic pathway of apoptosis by regulation of the tumor suppressor Death-Associated Protein Kinase 1 (DAPK1). *Thromb. Res.*, **2011**, *127*(2), 141-148.

[113] Tapper, W.; Hammond, V.; Gerty, S.; Ennis, S.; Simmonds, P.; Collins A; Prospective study of Outcomes in Sporadic versus Hereditary breast cancer (POSH) Steering Group, E., D. The influence of genetic variation in 30 selected genes on the clinical characteristics of early onset breast cancer. *Breast Cancer Res.*, **2008**, *10*(6), R108.

[114] de Fraipont, F.; Levallet, G.; Creveuil, C.; Bergot, E.; Beau-Faller, M.; Mounawar, M.; Richard, N.; Antoine, M.; Rouquette, I.;

Favrot, M.; Debieuvre, D.; Braun, D.; Westeel, V.; Quoix, E.; Brambilla, E.; Hainaut, P.; Moro-Sibilot, D.; Morin, F.; Milleron, B.; Zalcman, G. An Apoptosis Methylation Prognostic Signature for Early Lung Cancer in the IFCT-0002 Trial. *Clin. Cancer Res*., **2012**, *18*(10), 2976-2986.

[115] Yoo, H.; Byun, H.; Kim, B.; Lee, K.; Park, S.; Rho, S. DAPk1 inhibits NF-κB activation through TNF-α and INF-γ-induced apoptosis. *Cell Signal*., **2012**, *24*(7), 1471-1477.

Received: July 20, 2012 **Revised: September 26, 2012** **Accepted: ?????? 11, 2013**