Int. J. Data Mining and Bioinformatics, Vol.

Discovering breast cancer drug candidates from biomedical literature

Jiao Li and Xiaoyan Zhu

State Key Laboratory of Intelligent Technology and Systems,
Tsinghua National Laboratory
for Information Science and Technology,
Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, China
Fax: 86-10-62782266
E-mail: jiao-li04@mails.tsinghua.edu.cn
E-mail: zxy-dcs@tsinghua.edu.cn

Jake Yue Chen*

Indiana Center for Systems Biology and Personalized Medicine,
Indiana University,
School of Informatics,
Indianapolis, IN 46202, USA
Fax: 1-317-278-9201
E-mail: jakechen@iupui.edu
*Corresponding author

Abstract: We developed a new paradigm with the ultimate goal of enabling disease-specific drug candidate discovery with molecular-level evidences generated from literature and prior knowledge. We showed how to implement the paradigm by building a prototype literature-mining framework and performing drug–protein association mining for breast cancer drug discovery. In molecular pharmacology study of breast cancer, 79.2% of 729 enriched drugs in '*Organic Chemicals*' category were validated to be disease-related, and the remaining 20.8% were also investigated to be potential for future molecular therapeutics studies. '*Doxorubicin*', '*Etoposide*' and '*Paclitaxel*' were identified to have similar pharmacological profiles to treat breast cancer.

Keywords: biomedical text mining; structured data mining; drug identification; breast cancer.

Reference to this paper should be made as follows: Li, J., Zhu, X.Y. and Chen, J.Y. (2010) 'Discovering breast cancer drug candidates from biomedical literature', *Int. J. Data Mining and Bioinformatics*, Vol.

Biographical notes: Jiao Li is a PhD candidate at the Department of Computer Science and Technology, Tsinghua University. She received her BS in Computer Science from Northeastern University, China. Her research interests include biomedical literature mining, knowledge representation, data mining and disease-specific biomedical information integration.

Xiaoyan Zhu received her MS from Kobe University, Japan, in 1987, and her PhD from Nagoya Institute of Technology, Japan, in 1990. Currently, she is a Professor at the Department of Computer Science and Technology, Tsinghua University. She is the Principal Investigator of national keystone basic research program of China including the National Basic Research Program (973 program), National High Technology Research and Development Program of China (863 program) and National Natural Science Foundation of China (NSFC). She conducts active research in the area of pattern recognition, neural network, machine learning, natural language processing and information extraction.

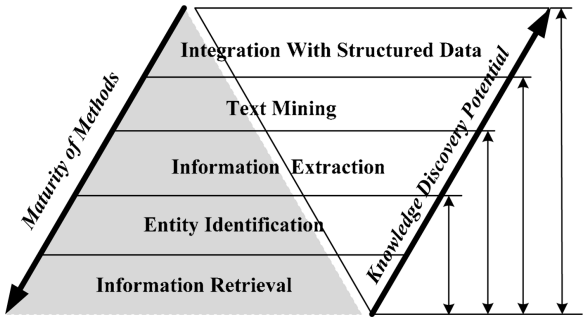
Jake Yue Chen received both his MS and PhD in Computer Science from the University of Minnesota in 1997 and 2001, respectively. He got his BS in Biochemistry and Molecular Biology from Peking University, China, in 1995. He founded the "Discovery Informatics and Computing Laboratory" at Indiana University – Purdue University Indianapolis (IUPUI) in early 2004. The research scope of the laboratory covers bioinformatics, scientific data management and data mining, functional genomics, proteomics, and systems and network biology. He is the founding director of Indiana Center for Systems Biology and Personalized Medicine (CSBPM), a senior member of the IEEE, and a senior member of the ACM.

1 Introduction

Recent advancement in the computational analysis and mining of biomedical literature databases can be categorised into five related topic areas: information retrieval (Schatz, 1997), entity identification (Leser and Hakenberg, 2005), information extraction (Vailaya et al., 2005), text mining (Cohen and Hersh, 2005) and integration of structured with textual data (Masys et al., 2001). The development of computational techniques that address research problems in these areas varies in their technical maturity and potential knowledge discovery impacts (Figure 1). In spite of technical challenges, many recent studies have begun to integrate techniques from a number of literature-mining topic areas to achieve maximal impacts on biomedical discovery from literature databases. For example, the EBIMED system combines information retrieval and entity recognitions to provide a complete picture of the retrieved results with different identified entities (EBIMed, 2007); the Textpresso system spans three topic areas by including information extraction to support individual sentence searching with concepts and relations between extracted objects (Müller et al., 2004); Srinivasan and Libbus (2004) developed a solution that spans four topic areas, to text mine implicit relationships from information extraction results; Korbel et al. (2005) further described a global literature-mining framework, in which keywords mined from literature may be integrated with a group of genes to reveal the phenotypic characteristics associated with these genes. Whereas tools based on these developments have helped biologists collect data automatically from large volumes of biological literature, how to turn them into biological knowledge discovery results still remain a challenging research question (Jensen et al., 2006). This is primarily because real-world biological problems are topic-specific (or, 'disease-specific' for biomedical research problems), therefore requiring text-mining systems to have not only decent recall but also high-precision – so that biologists do not have to sift through many 'likely' hypotheses before finding the 'golden nuggets'. Systems that attempt to integrate various

literature-mining components also tend to magnify type I errors accumulated from false positives (owing to the lack of high precision) within each component. These factors have made automated disease-specific knowledge discovery from biomedical literature far from a routine practice.

Figure 1 Current development and future trends of biomedical literature mining



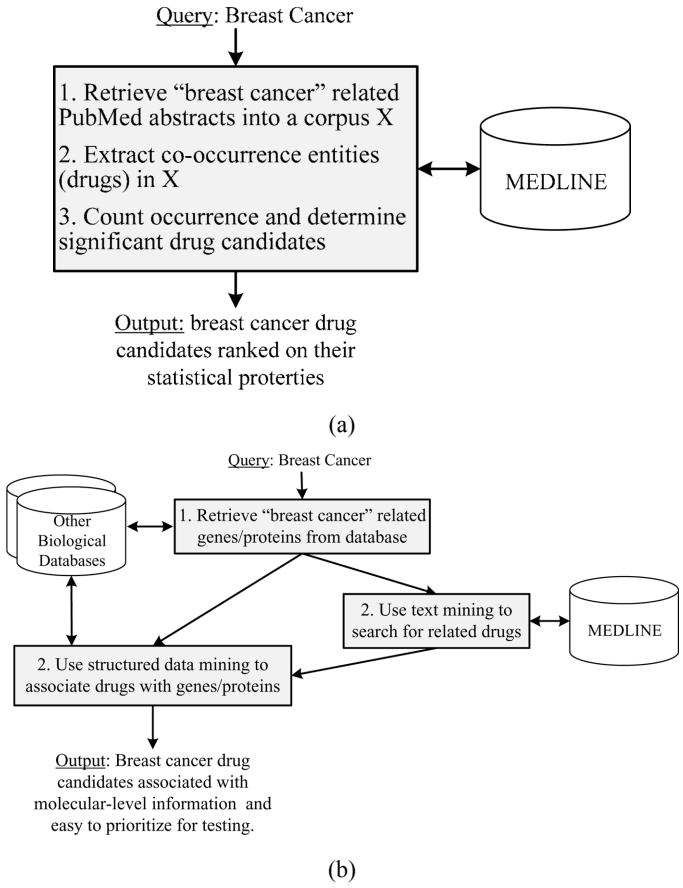
In this work, we present a novel literature-mining framework, with which we show how it can be used to enable novel molecular pharmacology studies of breast cancer. Unlike previous work that primarily concerns with one or part of the biomedical-literature-mining topic areas, our framework, driven by disease-specific biomedical knowledge discovery needs, integrated both textual- and structured-data-mining components. We developed and applied a suite of advanced statistical techniques to reduce false positives and random error. We applied this framework successfully to breast cancer drug candidate identification. Our results show several promising drug compounds worth investigating as candidates to treat breast cancer.

2 Comparison with existing work

Disease-related biomedical-literature-mining remains focused on how to improve the performance of information retrieval, entity recognition and information extractions. For example, G2D inferred logical chains of connections from disease names and ranked genes on the basis of a score that represented their likelihood of being associated with the query disease (Perez-Iratxeta et al., 2002). BITOLA extracted candidate genes that are indirectly connected to a given disease name (Hristovski et al., 2005). Tiffin et al. (2005) identified co-occurring disease name and tissue names in MEDLINE and linked the tissues to candidate disease genes. Srinivasan (2004) developed a solution to explore implicit relationships between pharmacology substances and diseases. Given disease name and user-specified terms, the above-mentioned biomedical-literature-mining techniques were capable of prioritising terms (e.g., genes, tissues and substances, etc.) with potential roles in the diseases (see Figure 2(a)). A common theme for these research activities is that they rely primarily on textual information retrieval engine to improve recall and precision, without taking advantage of prior knowledge from domain experts or molecular-level disease information stored in large online biological databases such as OMIM (Hamosh et al., 2005). Whereas these approaches are suitable for collecting disease-specific terminology information, it has largely become a 'hit-or-miss' to apply them to the discovery of molecular-level association information, e.g., disease-specific

drug–protein association information, which is one of the central significance in post-genome integrative and systems biology studies.

Figure 2 Comparison of two paradigms for disease-specific literature-mining systems: (a) a conventional paradigm and (b) our new paradigm



Compared with the existing work, we developed a new literature-mining paradigm with the ultimate goal of enabling disease-specific drug candidate discovery with molecular-level evidences generated from literature and prior knowledge. In Figure 2, we show a conventional paradigm used by most text-mining systems today (Figure 2(a)) and a new paradigm (Figure 2(b)) in breast cancer drug candidate identification.

Compared with the conventional paradigm, our new one has the following three distinct features. First, it constructs a list of disease-specific molecules such as proteins from prior knowledge. In post-genome biology, this prior knowledge may come from biological databases, which maintain large Omics experimental results (e.g., differentially expressed genes from microarray experiments comparing genes between disease samples and normal samples), or curated protein/gene databases for the given disease. Second, our paradigm incorporates not only the traditional textual-data-mining module that may be used to identify drugs from literature similar to the conventional text-mining system, but also a traditional data-mining module to associate drugs with molecules from

structured data sets. Third, the information processing flow is not linear: molecular-level information is used both for retrieving other types of molecular-level information and for building molecular association. Apparently, the end goal of the new paradigm is to challenge the traditional paradigm by identifying disease-related drugs and providing molecular-level association evidences. If shown feasible to implement, such paradigm could generate novel disease-related drug candidates previously not found in any given literature paper. Such conceptual paradigm, if successfully implemented, may represent a significant step forward, enabling literature-mining tools to generate hypothesis on which subsets of drugs are closely related owing to similarity in their molecular profiles.

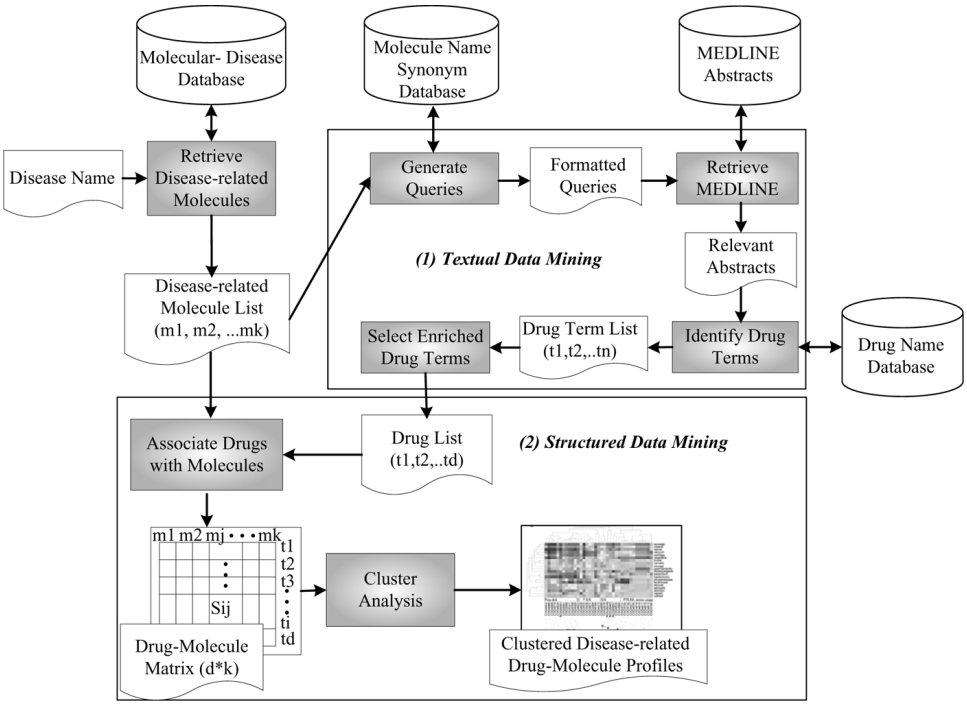
In this study, we build a prototype literature-mining system to show that such a new paradigm is feasible. We use breast cancer as our disease of interest, because it is a leading cause of cancer-related death in women worldwide with over 200,000 new diagnosed cases in the USA alone every year (Hinestrosa et al., 2007). Since current biomedical research are focused on building robust molecular biomarkers and finding drug treatments, we choose to focus on identifying breast cancer drug candidates whose potential therapeutic or toxicological effects are investigated at the molecular level where particular proteins are involved. With a list of breast cancer proteins from biological database, we want to study if we can acquire an association matrix between all breast-cancer-related drugs and proteins using the new paradigm. Our results next will indicate that several breast cancer drugs are found to share highly similar protein association profiles. Additionally, we found out that several drug candidates may be more promising in treating breast cancer than researchers realised before.

3 System framework

In this section, we describe how a computational-literature-mining prototype was developed to implement the new paradigm described early to generate drug candidates for breast cancer. In Figure 3, we show an overview of the framework for our implemented prototype.

The framework consists of two main modules. The first module, *Textual Data Mining*, takes the synonym-expanded disease-related molecule names constructed from biological database and outputs a list of drug terms that are significantly enriched in a collection of biomedical documents compared with the whole biomedical corpus from MEDLINE. The second module, *Structured Data Mining*, takes two inputs, disease-related molecule list and extracted drug list from the first module, and outputs a drug–molecule association matrix. In this module, statistical data cleaning and filtering and clustering analysis is applied to the association matrix in this drug candidate study for breast cancer. Other advanced data-mining techniques such as dimensionality reduction, attribute selection and classification may be used in other applications. The final clustered drug candidates contain novel knowledge (hypothesis), subject to validation. In the remainder of this section, we describe the framework methods in detail. In the next section, we report our experiment setting and result analysis.

Figure 3 A biomedical-literature-mining system framework to discover disease-specific drug candidates. It consists of two main modules as labelled: textual data mining and structured data mining



3.1 Construct breast cancer protein set

We built a curated list $\{m_1, m_2, \ldots, m_k\}$ of 214 breast cancer proteins, primarily based on genes from the Breast Cancer Gene Database (Baasiri et al., 1999) and some recent additions suggested by breast cancer experts (the list is available upon request). The genes from the database were mapped to UniProt Knowledgebase (UniProtKB) (Wu et al., 2006) identifiers using a biomedical gene/protein name thesaurus.

3.2 Retrieve MEDLINE abstracts and identify drug related to breast cancer proteins

The system maintains a comprehensive list of known name variants including acronyms, homonyms and synonyms by extending our early work (Li et al., 2005). It automatically generates an XML query statement with directives for synonyms and additional synonyms searching rules from user inputs (in our case, no additional user input rules were specified) for name recognitions. The following is an example query statement for the input, *BRCA1_HUMAN*:

$$< query > \#syn(\#uw6(Breast\ cancer\ susceptibility\ protein\ 1)$$
$$\#uw5(RING\ finger\ protein\ 53)\ BRCA1\ RNF53) < / query >$$

Here, the *#syn* directive instructs the query analyser to treat term or term expressions in the parameter set as synonyms, whereas the *#uwN* directive instructs the query analyser to match all term expression where the component terms are found within a neighbourhood of *N* adjacent words in any order. For example, the document (PMID = 14623252) with "*breast cancer type 1 susceptibility protein*" in its abstract will be retrieved by the above sample XML query, because the above phrase matches the "*#uw6 (Breast cancer susceptibility protein 1)*" query directive. In this study, we set *N* to be 1 plus the total count of terms within the directive.

From retrieved abstracts T_{BR} that are relevant to all the breast cancer proteins, the system automatically searches for terms if they appear in term categories of interests. The term categories are taken from '*Chemicals and Drugs*' in Medical Subject Headings (MeSH) (2007). We combine both term category dictionaries (*#syn* directives) and term category rules (*#uwN* directives) to identify drug terms $\{t_1, t_2, \ldots, t_n\}$ in T_{BR}.

3.3 Select significant breast cancer drugs

We use a term frequency statistical method described by Li and Chen (2009). This method makes use of term statistical distribution from the entire MEDLINE abstracts to calculate *p*-value of each term's significance in being observed in any collection of retrieved MEDLINE abstracts T_{BR}. It is a more sensitive method than the conventional *Term Frequency-Inverse Document Frequency* (*TF-IDF*) method in calculating the statistical significance of a term in a collection. The main reason for doing so is to control false positives among terms determined to be significantly enriched. For example, observing abnormally high usage frequency of a term from *TF-IDF* could lead to incorrect inclusion of the term as 'enriched', because the sampled document subset could be biased, and the term usage frequency could be intrinsically variable.

In this work, we first calculate a *p*-value for each drug term t_j in T_{BR} using methods described in Li and Chen (2009) and later derive its false discovery rate (FDR). Let the null hypothesis *Ho* be that document frequency of drug term t_j in T_{BR} comes from a random distribution T_{Random}. The *t-test* value Δ_j for drug term t_j can then be calculated as:

$$\Delta_j = (df(t_j \mid T_{BR}) - df(t_j \mid T_{\text{Random}})) \Big/ \sqrt{\frac{\text{Var}(t_j \mid T_{BR})}{N_{BR}} + \frac{\text{Var}(t_j \mid T_{\text{Random}})}{N_{\text{Random}}}} \tag{1}$$

where $\overline{df(t_j \mid T_{BR})}$ and $\overline{df(t_j \mid T_{\text{Random}})}$ are mean values of document frequencies of t_j in T_{BR} and T_{Random}, $\text{Var}(t_j|T_{BR})$ and $\text{Var}(t_j|T_{\text{Random}})$ are document frequency variances of t_j in T_{BR} and in T_{Random}, and $N_{BR} = |T_{BR}|$ and $N_{\text{Random}} = |T_{\text{Random}}|$ are collection sizes. *p*-value is computed as from two-sided tails $P(|Z| > |\Delta|)$ where $Z \sim N(0, 1)$:

$$p = P(|Z| > |\Delta|) = 2P(Z < -|\Delta|). \tag{2}$$

We used a standard multiple testing correction method (Benjamini and Hochberg, 1995) used in microarray analysis to convert *p*-value*s* from *t-test* to calculate a term's *FDRs*. In the end, we define all enriched terms meeting an empirically determined threshold (*document frequency* > 2 and *FDR* < 0.05) as significant drug terms $\{t_1, t_2, \ldots, t_d\}$.

3.4 Associate breast cancer drugs with proteins for cluster analysis

We assign an association score for each possible pair of significant drug terms $\{t_1, t_2, \ldots, t_d\}$ and breast cancer proteins $\{m_1, m_2, \ldots, m_k\}$, using a regularised log-odds function described in Korbel et al. (2005):

$$\text{Score}_{tm} = \ln(df_{tm} \times N + \lambda) - \ln(df_t \times df_m + \lambda). \quad (3)$$

Here, df_t and df_m are the total number of documents in which drug term and protein name are mentioned, respectively. df_{tm} is the total number of documents in which drug term t and protein name m are co-mentioned in the same document. N is the size of the entire MEDLINE collection. λ is a small constant ($\lambda = 1$ here) introduced to avoid out-of-bound errors if any of df_t, df_m, or df_{tm} values are 0. The resulting Score_{tm} is positive when the drug–protein pair is over-represented and negative when the drug–protein pair is under-represented. Since we do not care for this case study, we set all non-negative Score_{tm} to 0. Therefore, the higher the Score_{tm} is, the more significant the over-representation becomes.

In the structured data-mining module of the framework, we perform hierarchical clustering for significant breast cancer drugs using the Weighted Pair-Group Method with Half Square Euclidean Distance as the similarity measures. The similarity between drug t_a and t_b is calculated as follows:

$$\text{Sim}(t_a, t_b) = \frac{1}{2}\sum_{j=1}^{k}(\text{Score}_{t_a m_j} - \text{Score}_{t_b m_j})^2 \quad (4)$$

where $\text{Score}_{t_a m_j}$ and $\text{Score}_{t_b m_j}$ are cell values in drug–protein association matrix constructed by function (3). The similarity between proteins is also calculated by function (4). The final clustered attributes along the breast cancer drug dimension (vertical axis) are sorted by averaged values, increasing from top to bottom. The clustering task is both performed and visualised with the Spotfire DecisionSite Browser software (Dresen et al., 2003).

4 Experiment settings and result analysis

We report findings of our case study, using the prototype system described in the previous section. We analyse how related the retrieved drugs are to breast cancer, and whether the drug–protein association profiles help us gain new knowledge.

4.1 Categories of breast cancer drugs

In the textual data-mining module, the initial query contained 214 breast cancer proteins, without adding such suggestive terms as '*breast neoplasm*', '*breast lymphoma*', or '*breast cancer*' to test retrieval efficiency (in future routine practice, this could be added for improved precision). From 16,120,074 locally indexed MEDLINE records, breast-cancer-relevant abstracts were retrieved with queries corresponding to each protein and the results are unified to make a collection of 429,067 abstracts. We grouped the drug terms identified from these abstracts into 16 primary categories based on the subcategories of 'Chemicals and Drugs' in MeSH. In Table 1, we showed a distribution

of all the drug term counts (in the row '*#Total*') and the enriched counts (in the row '*#Enriched*') from the retrieved abstracts in the 16 'Chemicals and Drugs' MeSH term categories. The enriched drug terms, nearly half of the total in each category, were selected from all MeSH terms using the method described in Section 3.3. The quality and reliability of enriched drugs will be discussed in Section 4.2.

Table 1 Distribution of all drug terms and enriched ones identified from breast cancer retrieved abstracts

Category	*#Total*	*#Enriched*	*Category*	*#Total*	*#Enriched*
Inorganic chemicals	409	262	Lipids	252	151
Organic chemicals	1710	729	Amino acids, peptides, and proteins	2385	1573
Heterocyclic compounds	1088	422	Nucleic acids, nucleotides, and nucleosides	278	182
Polycyclic compounds	537	180	Complex mixtures	190	95
Macromolecular substances	160	101	Biological factors	651	399
Hormones, hormone substitutes, and hormone antagonists	187	112	Biomedical and dental materials	110	54
Enzymes and coenzymes	1195	658	Pharmaceutical preparations	57	39
Carbohydrates	335	167	Chemical actions and uses	438	292

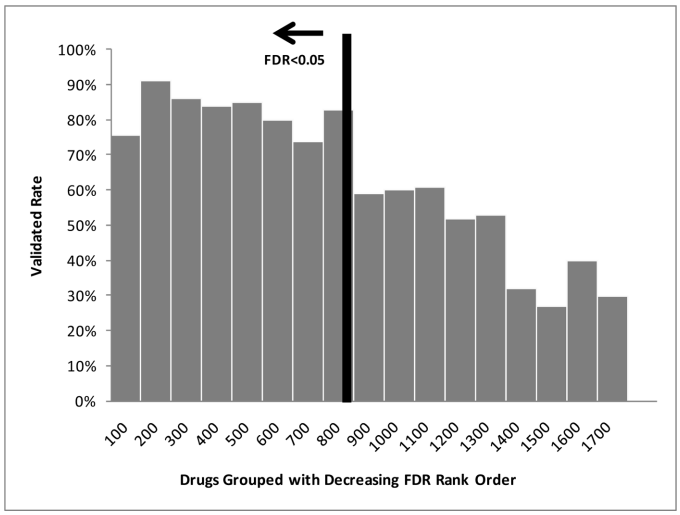
4.2 Validation of breast cancer drugs

To validate how well our method performed in breast cancer drug identification, we verified the results by building queries against MEDLINE abstracts in search of any publication, which contain any of the 1710 drugs in '*Organic Chemicals*' category and the term 'breast neoplasm' with all their term variants. Figure 4 shows the different validated rates under groups of drugs sorted by FDR descending rank orders. The validated rate decreases with the FDR threshold increasing, i.e., the more a drug is significantly enriched, the more it is likely relevant to breast cancer. There were 729/1710 or 57.3% drugs filtered out whereas the aggregate validated rate keeps 79.2%, if the FDR threshold is fixed at 0.05. In our 729 enriched drug list, these 79.2% co-cited drugs could be used by biomedical researchers new to the field to quickly gain an unbiased view which drugs are studied for breast cancer and on which protein context. A portion of the remaining 20.8% can be quite revealing for researchers to investigate.

Validation is built by querying against MEDLINE abstracts in search of any publication that contains any of the drugs and the term 'Breast Neoplasms' with all their term variants. *Validated Rate* is defined as the fraction of drugs that are validated under different groups of drugs sorted by FDR descending rank orders. i.e., *Validated Rate* = #(*Validated* ∩ *Drug_FDR*), where *#Drug_FDR* is the number of drugs grouped with their FDR rank order decreasing (here, 100 drugs per group) and *#Validated* is the number of drugs that are validated to be breast-cancer-related in each group.

To support the novelty of the remaining 20.8% drugs, we list 10 drugs extracted from the enriched breast cancer drugs without co-citing with 'breast cancer' in one MEDLINE abstract (see Table 2).

Figure 4 Histogram of different validated rates under drugs grouped with decreasing FDR rank orders



All the drugs in Table 2 were determined as the enriched one after calculating their *p*-values based on term statistical distribution from the entire MEDLINE. *False Discover Rate* was calculated using method of Section 3.3 "Select Significant Breast Cancer Drugs". *DF* is drug's document frequency in the breast-cancer-related corpus. Enriched drugs met the criteria of ($FDR < 0.05$ and $DF > 2$). The *Pharmacological Action* of each drug was annotated by reference to MeSH. Some drugs that occur frequently in the observed collection may not be more significantly enriched. For example, '*9,10-dimethyl-1,2-benzanthracene*' and '*Ethyl Chloride*' appeared in 402 and three abstracts in the breast-cancer-related corpus, and their FDR values were 0.008673 and 6.49E-12, respectively. Reasonably, these much more specific drugs are upgraded from the identified drug list.

All of the 10 drugs were validated to be novel breast cancer drug candidates, i.e., they are not co-cited with 'breast cancer' in on MEDLINE abstract. '*Ethyl Chloride*' is an organic solvent used in inhalation therapy as a rapid anaesthetic. It was hypothesised that exposure to such substances may cause breast cancer (Labrèche and Goldberg, 1997). '*Dizocilpine Maleate*' is considered for the wide variety of neurodegenerative conditions or disorders. '*Trichloroethylene*' is a highly volatile inhalation anaesthetic used mainly in short surgical procedures where light anaesthesia with good analgesia is required. Prolonged exposure to high concentrations of the vapour can lead to cardiotoxicity and neurological impairment. '*Amantadine*' is a well-known antiviral that is used in the prophylactic or symptomatic treatment of influenza A. It is also used as an antiparkinsonian agent, to treat extrapyramidal reactions, and for postherpetic neuralgia. These four drugs have pharmacological action of central nervous system agent, whereas breast cancer is the second most common cause of central nervous system metastases. Recent research described the clinicopathologic characteristics and

prognostic factors in breast cancer patients with central nervous system metastases (Altundag et al., 2007). '*Taurocholic Acid*' acts as a detergent to solubilise fats for absorption and is itself absorbed and is used as a cholagogue and cholerectic. '*Vitamin K1*' has antihaemorrhagic and prothrombogenic activity. '*Ethyl Methanesulphonate*' is an antineoplastic agent with alkylating properties, and also acts as a mutagen by damaging. '*9,10-dimethyl-1,2-benzanthracene*' is found in tobacco smoke that is a potent carcinogen. '*7,8-dihydro-7,8-dihydroxybenzo(a)pyrene 9,10-oxide*' derivatives have carcinogenic and mutagenic activity. They could be supported by the results indicating that smoking is associated with increased risk of breast cancer before age 50 years in BRCA1 and BRCA2 mutation carriers (Breast Cancer Family Registry et al., 2007).

Table 2 The enriched novel drug candidates in breast cancer studies

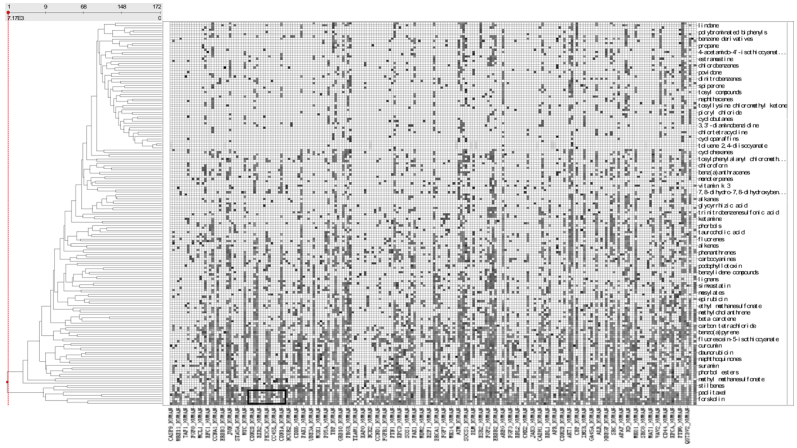
DRUG	*FDR*	*DF*	*Pharmacological action*
Ethyl chloride	6.49E-12	3	Sensory System Agents; Peripheral Nervous System Agents; Central Nervous System Depressants; Central Nervous System Agents; Anaesthetics, Local; Anaesthetics
Vitamin K1	6.70E-12	43	Vitamins; Haemostatics; Haematologic Agents; Growth Substances; Micronutrients; Fibrin Modulating Agents; Coagulants; Antifibrinolytic Agents
Ethyl methanesulphonate	1.34E-07	65	Mutagens; Noxae; Antineoplastic Agents, Alkylating; Antineoplastic Agents; Alkylating Agents
Taurocholic acid	6.78E-05	72	Gastrointestinal Agents; Surface-Active Agents; Detergents; Cholagogues and Choleretics
Dizocilpine maleate	0.001493	210	Neurotransmitter Agents; Neuroprotective Agents; Protective Agents; Excitatory Amino Acid Antagonists; Excitatory Amino Acid Agents; Central Nervous System Agents
Tosyllysine chloromethyl ketone	0.002944	38	Serine Proteinase Inhibitors; Protein Synthesis Inhibitors; Noxae; Protease Inhibitors; Enzyme Inhibitors; Alkylating Agents
Trichloroethylene	0.00671	25	Solvents; Central Nervous System Agents; Central Nervous System Depressants; Anaesthetics, Inhalation; Anaesthetics, General; Anaesthetics
7,8-dihydro-7, 8-dihydroxybenzo(a) pyrene 9,10-oxide	0.008521	65	Mutagens; Noxae; Carcinogens
9,10-dimethyl-1, 2-benzanthracene	0.008673	402	Noxae; Carcinogens
Amantadine	0.021045	28	Sensory System Agents; Neurotransmitter Agents; Peripheral Nervous System Agents; Dopamine Agents; Central Nervous System Agents; Antiviral Agents; Antiparkinson Agents; Anti-Dyskinesia Agents; Anti-Infective Agents; Analgesics, Non-Narcotic; Analgesics

These verification results directly affect how well we could extract new knowledge for exploring breast cancer drug candidates. Some of these enriched drugs may be good candidates for future pathological and molecular therapeutics studies. The enriched drugs also provide a confident list for subsequently built breast cancer molecular pharmacological profiles linking drugs with proteins.

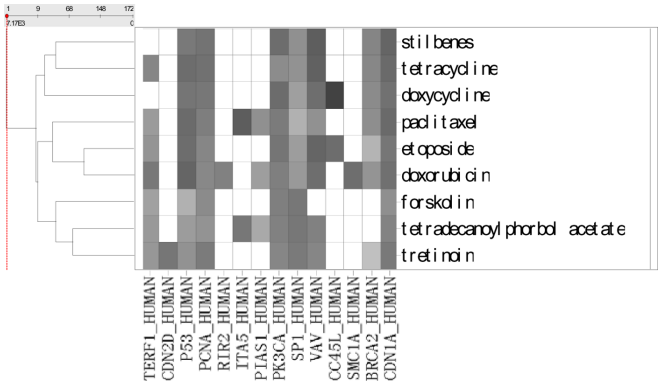
4.3 Cluster analysis of breast cancer drug–protein association

We applied a term category filter hydrocarbon drug compound from the MeSH category 'Chemicals and Drug' to obtain 172 drug entities for subsequent cluster studies (available upon request). We constructed an association matrix with the 172 drugs of as row, 214 breast cancer proteins as column and drug–protein association scores (defined in Section 3.4) as cell values in the 172×214 matrix. After hierarchical clustering (described in Section 3.4), we show the following results (Figure 5).

Figure 5 Hierarchical clustering result for breast cancer drugs: (a) global cluster results for the 172×214 matrix, where *x*-axis is for 214 breast cancer proteins and *y*-axis is for 172 hydrocarbon drugs and (b) zoomed view from boxed region in (a). Cell colour intensity is proportional association score (see online version for colours)



(a)



(b)

This matrix can help biomedical researchers compare molecular pharmacological profiles of different drugs studied in breast cancer and reveal available patterns among a group of functionally similar drugs. In Figure 5(a), we see distinctions between breast cancer more related drugs (rows at the bottom) and less related drugs (rows at the top). In Figure 5(b), several significant drugs in the context of breast cancer are apparently clustered together based on their protein association profiles. We could find interesting relationships between drugs that share similar molecular pharmacological profiles. For example, '*Doxorubicin*', '*Etoposide*' and '*Paclitaxel*' are clustered closely. All of them actually share similar chemical structure, with a cyclical hydrocarbons branch with carbon and hydrogen forming a closed ring. This is quite revealing because new breast cancer drugs could be developed on this shared chemical substructure. It is also revealing that although these drugs are predominantly known as anti-bacterial and antineoplastic agents, they may be studied for treating breast cancer, since the proteins that they were studied are strongly related to breast cancer. Another interesting observation is '*Losartan*', a drug compound for treating hypertension. Even though it is not associated with most of the breast cancer proteins significantly, its associations with some proteins such as '*P53*', '*GRB2*' and '*ERK2*' (general cancer-related genes) in the matrix suggest that it may be used as a general-purpose anti-cancer drug. We believe this association matrix contains many novel hypotheses subject to further testing by researchers in pharmaceutical industry R&D labs.

5 Conclusion and future work

In this work, we developed an integrated biomedical-literature-mining framework, with the ultimate goal of enabling disease-specific drug candidate discoveries. The framework successfully works in

- query-driven abstract retrievals and information extraction
- cluster analysis drug–protein association matrix in breast cancer, indicating promising results that can lead to many interesting subsequent biomedical research studies.

We believe several factors contributed to the successful application of our integrated biomedical-literature-mining framework. First, we fed the information retrieval engine with highly relevant molecule names instead of general high-level descriptive names. The specific molecular-level information built into the queries guaranteed to a certain degree that retrieved abstracts were specific to the inherent functional contexts of the molecules involved. Thus, it improves the precision of our information retrieval system. Second, we applied a suite of advanced statistical techniques, e.g., use of term frequency statistical method instead of the conventional *TF-IDF* methods to measure term frequency significance, use of FDR to select significant drugs and application of adjusted log-odds function to score drug–protein associations. The combination of these practical techniques made it possible to increase data processing efficiency and reduce error. Third, to accomplish our goals, we combined major advancements in both textual and structured data mining in this one, i.e., to perform MEDLINE abstract retrieval, to expand molecular names with thesaurus-driven entity recognition and query specification,

to extract disease-relevant drugs from retrieval results, to perform clustering analysis and to perform integrated data analysis and result validation.

The literature-mining framework presented in this work needs to be tested with additional biomedical domains besides breast cancer to create widespread implication in the field. However, we believe the framework is flexible, and can serve as a general guideline to many such ongoing case studies in the near future.

Acknowledgements

The work is in part supported by IUPUI Research Support Fund Grant (2006–2008), National Basic Research Program of China (973 Program) under No. 2007CB311003, National High Technology Research and Development Program of China (863 Program) under No. 2006AA02Z321, as well as NSFC under Grant No. 60572084 and No. 60621062. The authors also thank Scott H. Harrison for his critical review and his comments to help improve this manuscript.

References

Altundag, K, Bondy, M.L., Mirza, N.Q., Kau, S., Broglio, K., Hortobagyi, G.N. and Rivera, E. (2007) 'Clinicopathologic characteristics and prognostic factors in 420 metastatic breast cancer patients with central nervous system metastasis', *Cancer*, Vol. 110, No. 12, pp.2640–2647.

Baasiri, R.A., Glasser, S.R., Steffen, D.L. and Wheeler, D.A. (1999) 'The breast cancer gene database: a collaborative information resource', *Oncogene*, Vol. 18, No. 56, pp.7958–7965.

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society*, Vol. 57, No. 1, pp.289–300.

Breast Cancer Family Registry, Kathleen Cuningham Consortium for Research into Familial Breast Cancer and Ontario Cancer Genetics Network (2007) 'Smoking and risk of breast cancer in carriers of mutations in BRCA1 or BRCA2 aged less than 50 years', *Breast Cancer Research and Treatment*, Vol. 109, No. 1, pp.67–75.

Cohen, A.M. and Hersh, W.R. (2005) 'A survey of current work in biomedical text mining', *Briefs in Bioinformatics*, Vol. 6, No. 1, pp.57–71.

Dresen, I.M., Hüsing, J., Kruse, E., Boes, T. and Jöckel, K.H. (2003) 'Software packages for quantitative microarray-based gene expression analysis', *Current Pharmaceutical Biotechnology*, Vol. 4, No. 6, pp.417–437.

EBIMed (2007) www.ebi.ac.uk/Rebholz-srv/ebimed/

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. and McKusick, V.A. (2005) 'Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders', *Nucleic Acids Research*, Vol. 33, Database Issue, pp.D514–D517.

Hinestrosa, M.C., Dickersin, K., Klein, P., Mayer, M., Noss, K., Slamon, D., Sledge, G. and Visco, F.M. (2007) 'Shaping the future of biomarker research in breast cancer to ensure clinical relevance', *Nature Reviews Cancer*, Vol. 7, pp.309–315.

Hristovski, D., Peterlin, B., Mitchell, J.A. and Humphrey, S.M. (2005) 'Using literature-based discovery to identify disease candidate genes', *International Journal of Medical Informatics*, Vol. 74, Nos. 2–4, pp.289–298.

Jensen, L.J., Saric, J. and Bork, P. (2006) 'Literature mining for the biologist: from information retrieval to biological discovery', *Nature Reviews Genetics*, Vol. 7, pp.119–129.

Korbel, J.O., Doerks, T., Jensen, L.J., Perez-Iratxeta, C., Kaczanowski, S., Hooper, S.D., Andrade, M.A. and Bork, P. (2005) 'Systematic association of genes to phenotypes by genome and literature mining', *PLoS Biology*, Vol. 3, No. 5, p.e134.

Labrèche, F.P. and Goldberg, M.S. (1997) 'Exposure to organic solvents and breast cancer in women: a hypothesis', *American Journal of Industrial Medicine*, Vol. 32, pp.1–14.

Leser, U. and Hakenberg, J. (2005) 'What makes a gene name? Named entity recognition in the biomedical literature', *Brief Bioinform*, Vol. 6, No. 4, pp.357–369.

Li, H. and Chen, J.Y. (2009) 'Improved biomedical document retrieval system with PubMed term statistics an expansions', *International Journal of Computational Intelligence in Bioinformatics and Systems Biology*, Vol. 1, No. 1, pp.74–85.

Li, J., Zhang, X., Hao, Y. Huang, M. and Zhu, X. (2005) 'Learning domain-specific knowledge from context – THUIR at TREC2005 genomics track', *Proceedings of 14th Text Retrieval Conference (TREC2005)*, Gaithersburg, USA.

Masys, D.R., Welsh, J.B., Fink, J.L., Gribskov, M., Klacansky, I. and Corbeil, J. (2001) 'Use of keyword hierarchies to interpret gene expression patterns', *Bioinformatics*, Vol. 17, No. 4, pp.319–326.

Medical Subject Headings (MeSH) (2007) http://www.nlm.nih.gov/mesh/

Müller, H.M., Kenny, E.E. and Sternberg, P.W. (2004) 'Textpresso: an ontology-based information retrieval and extraction system for biological literature', *PLoS Biology*, Vol. 2, No. 11, p.e309.

Perez-Iratxeta, C., Bork, P. and Andrade, M.A. (2002) 'Association of genes to genetically inherited diseases using text mining', *Nature Genetics*, Vol. 31, No. 3, pp.316–319.

Schatz, B.R. (1997) 'Information retrieval in digital libraries: bringing search to the net', *Science*, Vol. 275, No. 5298, pp.327–334.

Srinivasan, P. (2004) 'Text mining: generating hypotheses from MEDLINE', *Journal of the American Society for Information Science and Technology*, Vol. 55, No. 5, pp.396–413.

Srinivasan, P. and Libbus, B. (2004) 'Mining MEDLINE for implicit links between dietary substances and diseases', *Bioinformatics*, Vol. 20, Supp. 1, pp.i290–i296.

Tiffin, N., Kelso, J.F., Powell, A.R., Pan, H., Bajic, V.B. and Hide, W.A. (2005) 'Integration of text- and data-mining using ontologies successfully selects disease gene candidates', *Nucleic Acids Research*, Vol. 33, No. 5, pp.1544–1552.

Vailaya, A., Bluvas, P., Kincaid, R., Kuchinsky, A., Creech, M. and Adler, A. (2005) 'An architecture for biological information extraction and representation', *Bioinformatics*, Vol. 21, No. 4, pp.430–438.

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'Donovan, C., Redaschi, N. and Suzek, B. (2006) 'The Universal Protein Resource (UniProt): an expanding universe of protein information', *Nucleic Acids Research*, Vol. 34, Database Issue, pp.D187–D191.