Direct Use of Information Extraction from Scientific Text for Modeling and Simulation in the Life Sciences

Martin Hofmann-Apitius, Erfan Younesi and Vinod Kasam
Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany

Abstract

Purpose:
To demonstrate how the information extracted from scientific text can be directly used in support of life science research projects. In modern digital-based research and academic libraries, librarians should be able to support data discovery and organization of digital entities in order to foster research projects effectively; thus we speculate that text mining and knowledge discovery tools could be of great assistance to librarians. Such tools simply enable librarians to overcome increasing complexity in the number as well as contents of scientific literature, especially in the emerging interdisciplinary fields of science. In this paper we present an example of how evidences extracted from scientific literature can be directly integrated into *in silico* disease models in support of drug discovery projects.

Design/methodology/approach:
The application of text-mining as well as knowledge discovery tools are explained in the form of a knowledge-based workflow for drug target candidate identification. Moreover, we propose an *in silico* experimentation framework for the enhancement of efficiency and productivity in the early steps of the drug discovery workflow.

Findings:
Our *in silico* experimentation workflow has been successfully applied to searching for hit and lead compounds in the World-wide *In Silico* Docking On Malaria (WISDOM) project and to finding novel inhibitor candidates.

Practical implications:
Direct extraction of biological information from text will ease the task of librarians in managing digital objects and supporting research projects. We expect that textual data will play an increasingly important role in evidence-based approaches taken by biomedical and translational researchers.

Originality / value:
Our proposed approach provides a practical example for the direct integration of text- and knowledge-based data into life science research projects, with the emphasis on its application by academic and research libraries in support of scientific projects.

Paper type:
Conceptual paper

Keywords:
Text mining, information extraction, *in silico* experiment, drug discovery, grid computing

1. Introduction

The Life Sciences (biology, biochemistry, medicine) are still dominated by empirical observations. Because of this empirical nature of the life sciences there is a flood of descriptive publications in this domain. Besides a remarkable increase in the complexity of the scientific content of life science publications (e.g. observations that cross the borders of traditional disciplines, indicated by new journals with names such as NATURE Chemical Biology), the number of journals is also growing fast. Approximately 13,000 biomedical journals are being published currently throughout the world, among which more than 5000 are currently indexed for MEDLINE in the fields of biomedicine and life sciences (www.nlm.hin.gov/pubs/factsheets). Moreover, about 120 new journals are added to MEDLINE every year (Kotzin, 2005). This increasing volume of information poses a great challenge to life scientists to search, retrieve and extract relevant data in an efficient and reliable manner. In response to this challenge, automated methods for information retrieval and information extraction ("text mining") have been developed and continuously improved. These technologies have recently reached a degree of maturity that enhances the searchability of traditional information retrieval systems through different techniques such as query refinement, semantic searching, document clustering and categorization, and summarization (Mack and Hehenberg, 2002). However, the indispensable part of information retrieval (IR) systems is information extraction (IE), which is intended to identify and extract specific biological terms (named entity recognition, NER) and their relationships automatically. IE techniques are evolving at a faster pace recently, making use of the rapid development of semantic annotations and ontologies, which help to classify mentions of text entities and enable true semantic search by mapping named entities to classes of entities (e.g. the named entity "Alzheimer" belongs to the class "diseases"). Unstructured information sources such as scientific text are rich in useful information on e.g. diseases and their molecular etiology. Such information is often represented in text by associations among similar or different biological entities (i.e. genes, proteins, drugs, allelic variants, etc). In addition to their capability to retrieve and extract "direct relationships" among biological entities (e.g. published facts), text-mining techniques can be leveraged to detect 'invisible' patterns or 'indirect' associations among different entity types. For example, if protein A interacts with protein B and protein B interacts with protein C, it can be inferred that protein A might also interact with protein C in a complex. Therefore, text mining can be used to enhance data mining capabilities. As such, it would be interesting to mine indirect associations, for example, between drug-drug, drug-pathway, drug-marker, and drug-clinical outcome information. In this paper, we explain how the information extracted from scientific text can be directly incorporated into the process of modeling pathogenesis and simulation of *in silico* experiments using computational tools and high-performance computing infrastructures.

2. Technologies used for information extraction from text and images

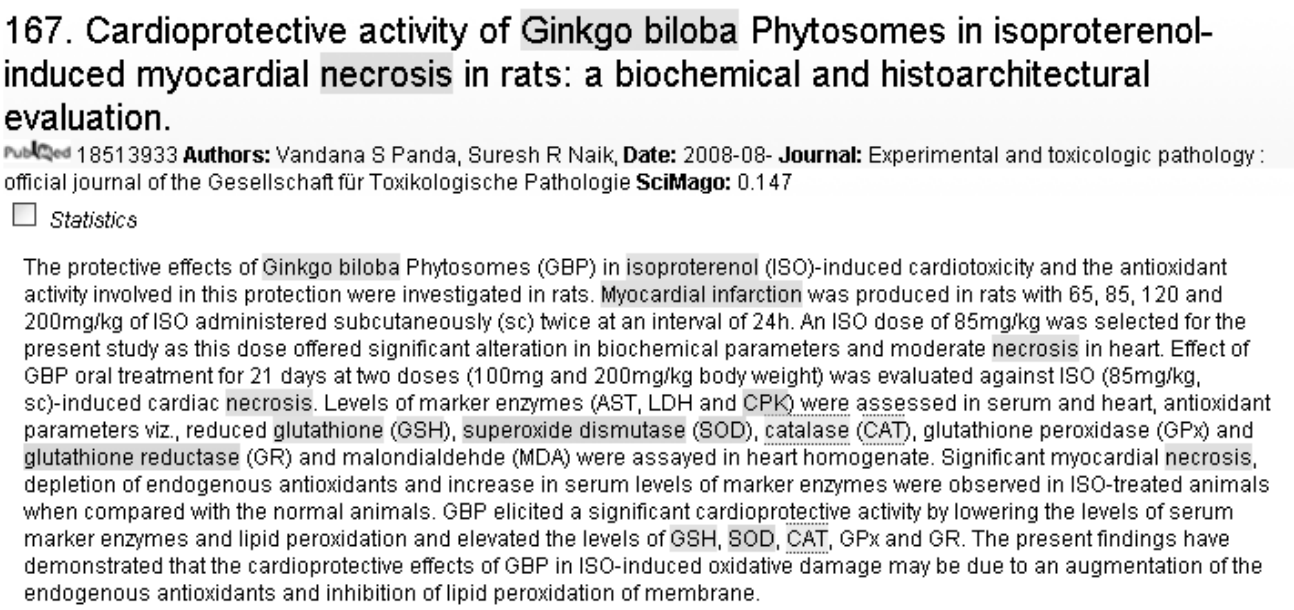
Technologies for Information extraction (IE) which are currently widely used are based on two approaches: natural language processing (NLP) and statistical methods, namely co-occurrence of entities (Jensen *et al.,* 2006). Using NLP methods, biomedical information within free text is mined through part-of-speech (POS) taggers or part-of-speech stemmers. While the first method labels each word in the sentence according to its grammatical position, the second one recognizes the morphological root of the word (stemming).
Since information extraction deals with the semantic structure of the text, the first step toward IE is to identify and tag biological entities in the text; this process, which is called "named entity recognition" (NER), is an active area of research due to the increasing complexity of biomedical language and vocabulary (Krallinger *et al.,* 2005). Rule-based methods, dictionary-based approaches or a combination of these two techniques are often used for information extraction from the text. Table 1 lists different methods which are currently used for IE purposes in the life sciences domain.

Table 1. List of different methods that are applied to Information Extraction

IE methods	Approach
Rule-based pattern extraction	Learned regular expression patterns
Sequence tagging by machine learning	Probabilistic sequence models (HMM*, CRF**), induced classifiers
Dictionary-based pattern matching	Named entity recognition

* Hidden Markov Models (Rabiner, 1989)
** Conditional Random Fields (Lafferty *et al.,* 2001)

A prerequisite for IE is entity recognition which is considered one of the most challenging areas in text mining, mainly due to the lack of standard naming (Jensen *et al.,* 2006). Abbreviations and synonyms, which represent biological named entities in the text should be distinguished from the background; for instance, the gene name "AR" which stands for androgen receptor should be distinguished from the acronym for "Arkansas state". Therefore, recent systems are supported by ontologies or dictionaries containing a comprehensive list of synonyms in order to reduce the number of false positives; for example, ProMiner developed at Fraunhofer Institute for Algorithms and Scientific Computing, SCAI, (Hanisch *et al.,* 2005) is a rule-based system that is supported by regularly updated organism-specific dictionaries. The system resolves the ambiguities using context information and is aware of acronyms (Figure 1).



167. Cardioprotective activity of Ginkgo biloba Phytosomes in isoproterenol-induced myocardial necrosis in rats: a biochemical and histoarchitectural evaluation.

PubMed 18513933 **Authors:** Vandana S Panda, Suresh R Naik, **Date:** 2008-08- **Journal:** Experimental and toxicologic pathology : official journal of the Gesellschaft für Toxikologische Pathologie **SciMago:** 0.147

☐ *Statistics*

The protective effects of Ginkgo biloba Phytosomes (GBP) in isoproterenol (ISO)-induced cardiotoxicity and the antioxidant activity involved in this protection were investigated in rats. Myocardial infarction was produced in rats with 65, 85, 120 and 200mg/kg of ISO administered subcutaneously (sc) twice at an interval of 24h. An ISO dose of 85mg/kg was selected for the present study as this dose offered significant alteration in biochemical parameters and moderate necrosis in heart. Effect of GBP oral treatment for 21 days at two doses (100mg and 200mg/kg body weight) was evaluated against ISO (85mg/kg, sc)-induced cardiac necrosis. Levels of marker enzymes (AST, LDH and CPK) were assessed in serum and heart, antioxidant parameters viz., reduced glutathione (GSH), superoxide dismutase (SOD), catalase (CAT), glutathione peroxidase (GPx) and glutathione reductase (GR) and malondialdehde (MDA) were assayed in heart homogenate. Significant myocardial necrosis, depletion of endogenous antioxidants and increase in serum levels of marker enzymes were observed in ISO-treated animals when compared with the normal animals. GBP elicited a significant cardioprotective activity by lowering the levels of serum marker enzymes and lipid peroxidation and elevated the levels of GSH, SOD, CAT, GPx and GR. The present findings have demonstrated that the cardioprotective effects of GBP in ISO-induced oxidative damage may be due to an augmentation of the endogenous antioxidants and inhibition of lipid peroxidation of membrane.

Figure 1. Visualization of ProMiner performance on recognition of biological named entities. In this example, the word “CAT” has been correctly detected as gene name for Catalase; genes/proteins, drugs, and disease are also highlighted by gray

As well as free text, images constitute another source of relevant information, which is frequently found in publications, especially in the chemical and biochemical domain. The above-mentioned text mining technologies have contributed significantly to the progress of recognition and extraction of chemical named entities from text and image captions, but the big challenge with chemical structure depictions is how to convert a structural image to a computer readable structure representation format; such structural data can be stored in searchable databases and used for drug discovery purposes (Banville, 2006). The first attempts at automated extraction of chemical structure information from images and their conversion into computer readable chemical structure representation formats appeared in the 1990s (McDaniel and Balmuth, 1992). A commercial tool for extraction of chemical data from literature, CLiDE, (standing for “chemical literature data extraction”) was developed in the middle of the 1990s (Ibison *et al.,* 1993). Very recently, an advanced tool for chemical structure mining (chemoCR^{TM}) has been developed at Fraunhofer Institute SCAI. chemoCR combines pattern recognition techniques with supervised machine-learning concepts and a chemical expert system in order to identify the most significant semantic entities (e.g. chiral bonds, super atoms, reaction arrows, etc.) from chemical depictions. The system is still being improved, but the current version is now “production ready”, meaning that chemical structures can be directly used as input into *in silico* experimentation such as virtual screening.

3. Direct application of extracted data from text to *in silico* experiments

Given such considerable advancements in information extraction from text and images, one might ask how these data extracted from the literature can be used to enhance our understanding and knowledge in the biomedical domain. In their review, Krallinger *et al.* (2005) point out four applications for text mining, namely functional annotation of genes and proteins, extraction of subcellular localizations, statistical analysis of gene expression articles, and prediction of protein-protein interactions. Among these applications, the statistical approaches have been extended to annotating the content of expression databases; for example, very recently Ruau *et al.* (2008) used ProMiner (Hanisch *et al.,* 2005) to annotate data entries (biological sample information) in the Gene Expression Omnibus (GEO) microarray repository by employing text-mining and expression profile correlation. In this way, the annotation process could be automated.

In addition to the usage scenarios mentioned by Krallinger *et al.* (2005), we foresee that extracted data from scientific publications can be directly used in *in silico* experiments. An *in silico* experiment has been defined as "a procedure that uses computer-based information repositories and computational analysis to test a hypothesis, derive a summary, search for patterns, or demonstrate a known fact" (Foster and Kesselman, 1999). In other words, an *in silico* experiment involves the use of local and remote resources to test a hypothesis (Stevens *et al.,* 2003). Since biological systems comprise of dynamic interactions between non-linear processes on tempo-spatial scales, their analysis and modeling requires integration of all relevant information at multiple levels from molecular and cellular to organ levels. For this reason, life scientists frequently need to collect information from different databases such as EntrezGene, SwissProt, or PDB (Protein Data Bank), and use them in combination in order to be able to test their hypotheses in the *in silico* environment before they proceed to the more expensive and time-consuming experimental lab work. Therefore, computer models and simulation environments provide the researchers with a convenient test ground to go through the process of "trial and error" and further optimize and validate the outcome of their experiments before proceeding to the real experimental settings in the molecular biology laboratory. In the following, we demonstrate how literature-based information is directly used in *in silico* modeling and simulation experiments.

3.1 *In silico* Network-based modeling of complex diseases

Many human diseases (over 1500) have been found to result from a defect in the function of a single gene; for example sickle-cell anaemia is a blood disorder which is characterized by abnormal sickle shape of red blood cells because of a mutation in the haemoglobin gene. Such so-called "Mendelian" diseases occur rarely and their transmission follows a characteristic pattern (e.g. dominant, recessive, sex-linked). However, there are many other

diseases that are more “common” in the human population and their inheritance follows a familial pattern; such diseases are often referred to as “complex diseases” because they are not the result of simple Mendelian inheritance (Botstein and Risch, 2003); instead they are likely to arise from mutations in more than one gene or different mutations in the same gene (Goh *et al.,* 2007). Usually an unknown number of multiple defected genes are involved, which are also interacting with environmental factors and lead to the manifestation of such diseases as coronary heart disease, hypertension, diabetes, obesity, various cancers, and neurodegenerative diseases (Motulsky, 2006).

The polygenic nature of complex diseases (contribution of many mutated genes with low effect) has a great impact on the underlying cellular network at different molecular levels from gene expression to proteomic and metabolic levels. In other words, for us to be able to understand the biological mechanism(s) underlying complex diseases, it is necessary to consider the contribution of all possible defected genes and their products in a network of dependencies. Nowadays, high-throughput technologies have made it possible to look at the disease state from a global or system view and have produced a large amount of data at each molecular level. For instance, genome chip and microarray technology now allows us to study the activity of large numbers of genes simultaneously and to create a global picture of cellular function under different conditions (disease vs. healthy samples).

Integration of all such data into comprehensible models using computational tools allows us to understand the biological complexity behind complex diseases by simulating the behaviour of cells under disease conditions in a virtual environment. *In silico* modeling provides a suitable framework for the integration of high-dimensional data across different biological domains which can be used for hypothesis generation and prediction; for example, in a cellular interaction network proper intervention points can be hypothesized as drug target candidates and modulating these points *in silico* may predict the clinical readout at the phenotypic level (Butcher *et al.,* 2004). Recent application of network theory to the biological field has laid down the foundation of a model framework known as “integrative functional informatics” or “integrative bioinformatics” (Figure 2).

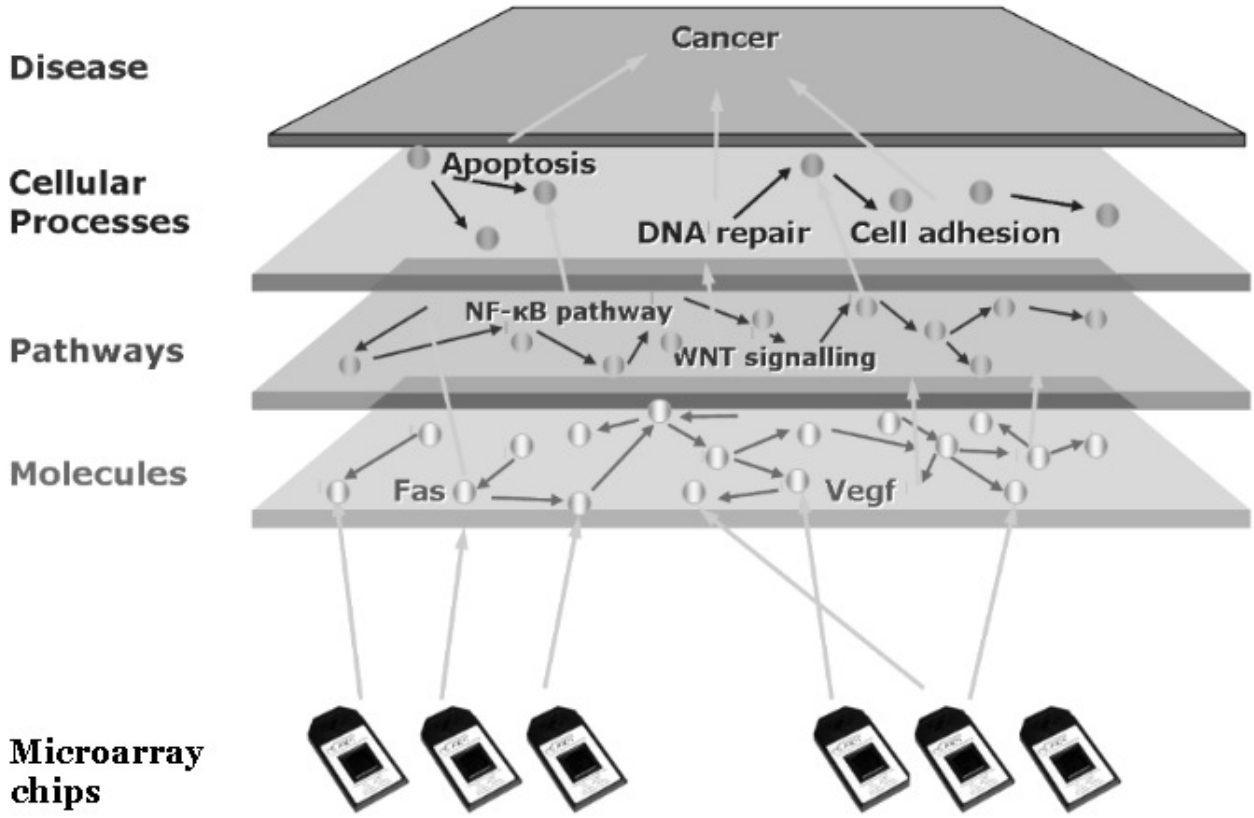


Figure 2. A systems biology view at the pathogenesis mechanisms in cancer. The aim of integrative bioinformatics is to aggregate the biological data at different levels and derive a model that is an indicator of underlying cellular interactions under disease state.

Network-based models have been successfully constructed and employed to tackle the complexity of common diseases from different angles, including identification of disease-modifying proteins in ataxia (Lim *et al.,* 2006), finding novel drug targets for prostate adenocarcinoma (Ergun *et al.,* 2007), analysis of allergic response in asthma (Lu *et al.,* 2007), and network-based classification of breast cancer metastasis (Chuang *et al.,* 2007). Such networks represent the reduced form of complex interactions among cellular components so that each node stands for a molecular component and two nodes are connected with a link if they interact. The nodes and links together form an interaction network, which is translated into the mathematical language of graphs (Barabasi and Oltvai, 2004). Networks serve as models for the integration of cellular information as well as the generation of a predictive hypothesis about the behavior of a biological system; for example, if changes in gene activity can be mapped to changes of corresponding secreted proteins in the blood, perhaps diagnosis of the disease could be easily done by testing the amount of such proteins in the blood (Hood *et al.,* 2004).

Evidence extraction from literature is often the preliminary step in formulating a network hypothesis for disease modeling. For instance, Pujana *et al.* (2007) first constructed a

literature-based network of human Protein-Protein Interaction (PPI) using four reference genes and then projected gene co-expression profiles onto this network; they enriched the resultant networks with additional information from human and other model organisms and successfully identified and experimentally validated a new component of this susceptibility network.

Each individual has a unique genetic makeup and this uniqueness accounts for the phenotypic variations observed among individuals in the human population (Kruglyak and Nickerson, 2001). These variations in genetic composition (also called “genetic polymorphisms”) may have a great impact on disease susceptibility. Such polymorphisms constitute the genetic component of complex human diseases but they are thought to trigger the process of pathogenesis under the influence of environmental factors. To find out which genetic variants increase disease risk, all the variations across the human genome are genotyped and those variants which are quantitatively recognized as risk factors are determined. Identification of susceptibility genes is a progressing field of research, as it is expected that many susceptibility variants will be discovered in the future (Iles, 2008).

In an attempt to identify the most important susceptibility players in breast cancer and to explore their relationships with other known susceptibility mutations in humans, our group at Fraunhofer SCAI used a network-based approach to test the hypothesis that single-point errors in the genetic code of multiple proteins lead to an increased level of susceptibility to breast cancer and that the degree of susceptibility depends on the position and function of each protein in the entangled network of cellular interactions. For this purpose, a human PPI network, relevant to breast cancer, was constructed and the susceptibility dataset drawn from 1140 patients with breast cancer (Hunter *et al.,* 2007) was mapped onto this network. This network was topologically characterized and also compared to the randomized version. Topological and functional analyses of this network identified 13 significant genes which might play a central role in conferring susceptibility to the development of breast tumors. In parallel, we employed SCAIview – a knowledge discovery tool developed at Fraunhofer SCAI (Friedrich *et al.,* 2008) - to reconstruct a literature-based network of gene co-citations relevant to breast cancer. To explore novel susceptibility associations between our 13 genes and other known ones, we overlaid the co-citation network on the PPI network and found overall 23 novel associations from which 7 associations could be directly or indirectly validated by the literature. Our findings are consistent with the fact that many susceptibility genes have not yet been discovered due to the low heritability of complex traits as well as the underpowered statistical methods used in linkage analyses (Hirschhorn and Daly, 2005). This example shows that the information embedded in free text can be used for more sophisticated purposes than simply extraction of biological entities. Enrichment of molecular

network analyses with text-mining data not only increases the added value of the analysis, but also strengthens the validity and interpretability of the results.

3.2 *In silico* experimental environment for high-throughput screening

With the advent of high-throughput technologies, researchers are now confronted with massive amounts of biological data, which have to be analyzed and interpreted with the help of bioinformatics applications. By integrating different biological datasets, life scientists are able to study the biological system as a whole and thus systems biology approaches are becoming more popular in the course of routine research activities. However, pattern detection, modeling and simulation of the biological system, and hypothesis testing are prerequisite steps in the cycle of systems biology approaches (Figure 3). This strategy has already been adopted by the pharma industry and academia for drug discovery purposes. One limiting factor in adopting this strategy is the demand for high-performance computational capacities. Hence, the concept of the "virtual laboratory" has been introduced, in which computational distributed resources are used as an electronic workspace for drug target identification, selection, and validation (Rauwerda *et al.,* 2006). Grid technology provides a computational backbone for this purpose (Konagaya, 2006).

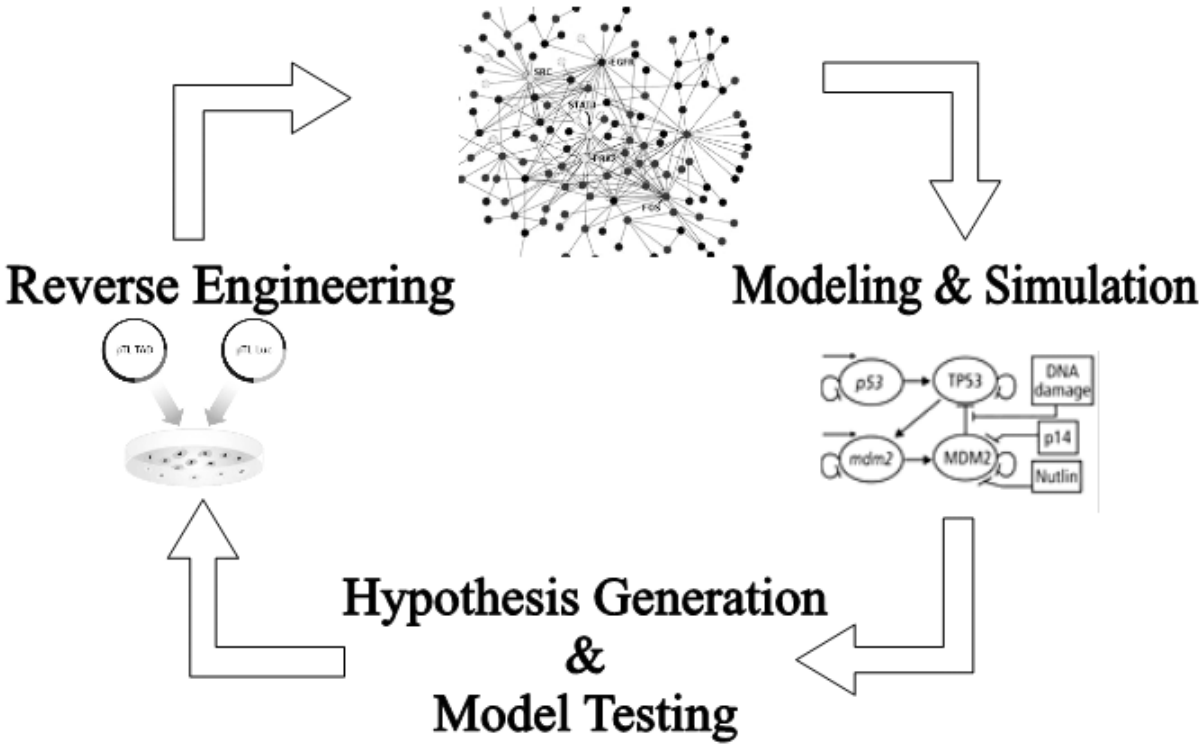


Figure 3. The cycle of systems biology approaches to disease modeling and prediction.

High-throughput virtual screening by molecular docking is an example of an *in silico* experiment which replaces the high-cost procedure of real HTS (High-Throughput Screening) in experimental laboratories and makes it possible to perform screening of millions of compounds on interesting target proteins in a reliable, rapid, and cost-effective manner (Jacq

et al., 2006). This approach has been successfully applied to searching for hit and lead compounds in the World-wide *In Silico* Docking On Malaria (WISDOM) project (Jacq *et al*., 2008).

The WISDOM project encompasses a collaborative framework, which has been established between bioinformaticians, biochemists, pharmaceutical chemists, biologists and grid computing experts in order to produce and make selected lists of potential inhibitors available. WISDOM-I, the first large scale deployment of the molecular docking application on EGEE (Enabling Grids for E-sciencE) [i], which took place from August 2005 to September 2005, has seen 42 million dockings, which is equivalent to 80 years of CPU time. Virtual screening of 500,000 chemical compounds was performed using FlexX software against different plasmepsins (aspartic protease implicated in haemoglobin degradation). On the biological front, three scaffolds were identified, one of which is the guanidino scaffold, which is likely to be novel as they have not been reported as plasmepsin inhibitors before. Experimental results proved that the compounds selected from WISDOM-I function as sub-micromolar inhibitors against plasmepsin (Kasam *et al.,* 2007; Jacq *et al.,* 2008). The complete workflow employed in the WISDOM project is shown in Figure 4.

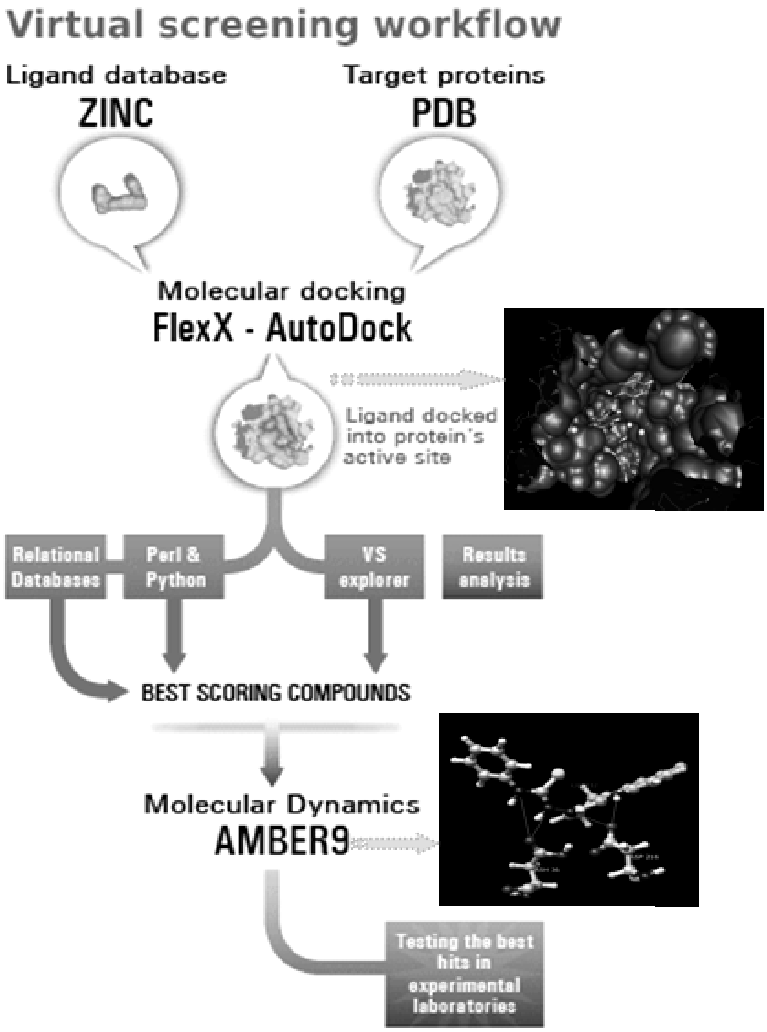


Figure 4. Description of the *in silico* experimentation workflow deployed in the WISDOM project for large scale screening of compounds against Malaria.

With the success achieved by the WISDOM-I project on both the computational and biological sides, several scientific groups around the world proposed targets implicated in malaria, which led to the second assault on malaria, WISDOM-II (Kasam *et al.,* 2007). The target portfolio was broadened, and the ZINC database (4.3 million chemical compounds) was screened against four different targets implicated in malaria. Over the course of 90 days during the winter of 2006, 140 million dockings were recorded, which is equivalent to 413 years of CPU time, representing an average throughput of almost 80,000 dockings per hour. This was made possible by the availability of thousands of CPUs through different infrastructures worldwide. Experimental testing of the compounds finally selected against all the targets is ongoing.

4. Linking disease modeling to grid-based target identification

Normally a research process starts with an exploration of a scientific domain by collecting relevant data, information, and previous knowledge, which are often hidden in scientific publications. Accordingly, referring to the scientific literature is usually the first step towards selection of drug targets and validation processes (Whittaker, 2004) because it provides a valid and proper framework for drug target identification purposes. When merged with network-based disease models, the information extracted from the text enhances confidence about the drugability of the candidate target. Moreover, it would be possible to generate informative profiles for each candidate target using information extracted from the text; i.e. literature-based annotation of target nodes on the network model of disease provides enormous insight about drug candidate efficacy and toxicity. Such profiles will be of high value for ranking or prioritizing target candidates. Another potential application field for this strategy is the emerging phenomenon of Polypharmacology in which the drugability of a specific ligand against multiple targets (rather than a single target) is assessed for treatment of polygenic complex diseases (Hopkins, 2008).

To this end, we have devised a knowledge-based workflow for target candidate identification, which incorporates the information extracted from the text directly into the network-based disease model (Figure 5). In this workflow, information retrieval is performed on PubMed abstracts using the user's search query in a context-sensitive manner. Information extraction is accomplished on a relevant subcorpus by a rule-based system that employs a machine-learning technique and resolves the ambiguity problem by using regularly updated organism-specific dictionaries (Hanisch *et al.,* 2005). The system returns the results according to statistical ranking of entities found, based on Kullback-Leibler divergence (relative entropy), meaning that the more relevant entities (e.g. gene names) appear in the top of the ranking list. This system is able to extract co-mentioned biological entities (e.g. gene-gene, protein-protein) and export them as a co-occurrence network together with the corresponding frequencies of co-mentions in the literature. These frequencies can be later used as the weight of edges in the co-occurrence network for filtering purposes. This co-occurrence network passes through the next module which compiles a protein-protein interaction (PPI) network from manually curated databases such as DIP (Datase of Interacting Proteins), BIND (Biomolecular Interaction Network Database), HPRD (Human Protein Reference Database), etc. Since the PPI network is constructed using expert-curated protein interaction data from databases, it provides a well-defined backbone for mapping the co-occurrence network and exploring potential novel associations suggested by the text mining approach. The output of the workflow is a network model which consists of both curated and text-based information and can be further enriched by different types of biological data adopted from molecular databases (e.g. gene expression values) or from the text itself (e.g. pathological or

clinical context). This network model is then subjected to statistical analysis to identify the key biological elements correlated with the pathogenesis mechanism.

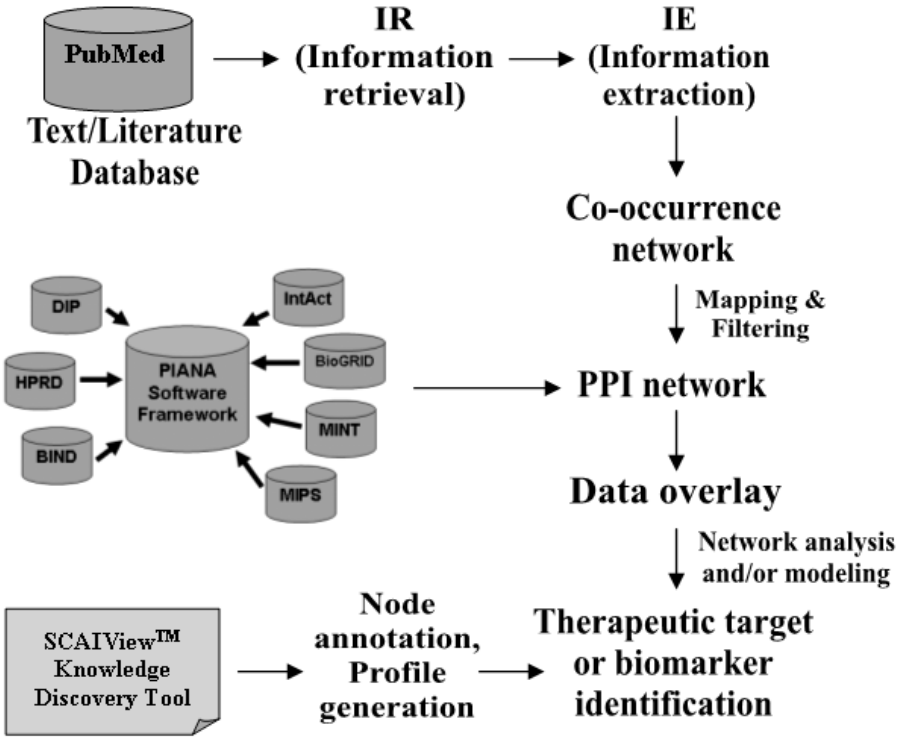


Figure 5. A knowledge-based workflow for target candidate identification. The information from the text is incorporated in the form of co-occurrences and annotations which are directly extracted by the SCAIView knowledge discovery tool.

The advantage of the text-mining data over expert curated data is that text-mining data are extracted from up-to-date information in the literature, thus increasing the chances of uncovering novel associations. Whereas manually (expert) curated data represent well-established knowledge (which is often represented in cartoon-like schemata), text-mining derived knowledge seems to be more suited to fostering the generation of novel hypotheses. The results from the above workflow are hypothetical suggestions and need to be tested in an *in silico* environment using simulation techniques; by this means, it can be ensured that the most promising target candidate will be selected to pass through the next expensive steps of the drug discovery pipeline. Automatic ligand-target dockings on high-performance grid computing infrastructures can help us effectively for this purpose: a library of numerous drug-like molecules is docked against candidate targets and consequently docking properties for all ligand-target combinations can be computed on a grid-enabled high-performance architecture in a very time- and computational-efficient manner. The most promising combinations are then selected and directed towards the next steps of the drug discovery pipeline (Figure 6).

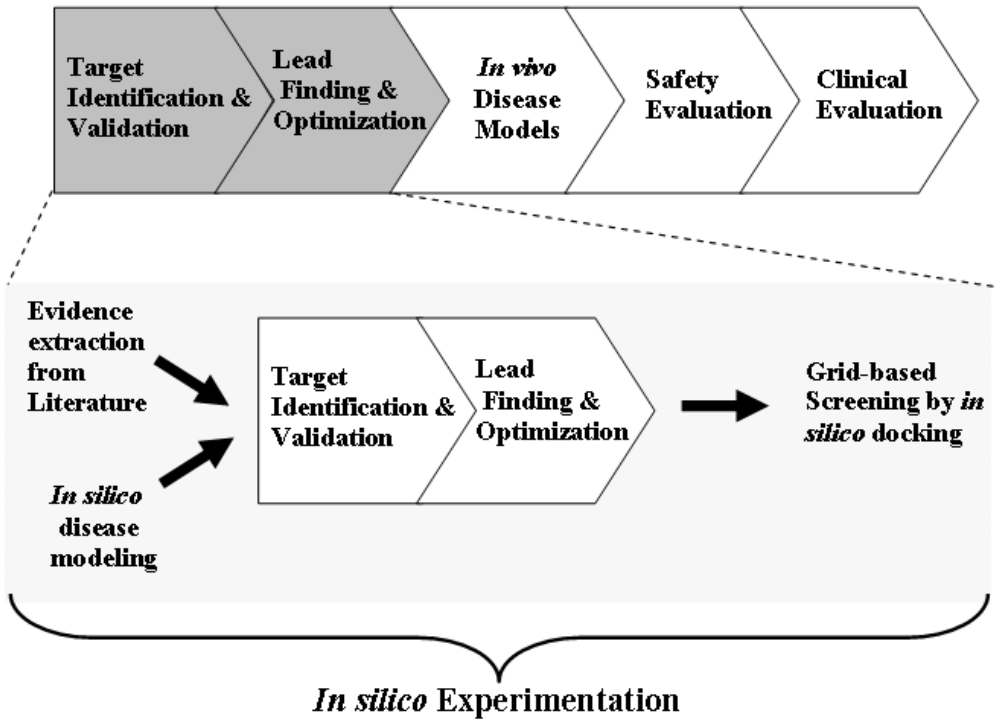


Figure 6. Proposed *in silico* experimentation framework for the enhancement of efficiency and productivity in the early steps of the drug discovery workflow.

In conclusion, the embedding of text and data mining methodologies in the architecture of *in silico* experimentation environments not only complements the experimental data but also enhances the predictive power of the resultant semantic model for the disease in question. Currently, drug development and productivity is facing a high rate of failure (as much as 30%) due to a lack of efficacy and clinical safety (Kola and Landis, 2004). Integration of literature evidence into *in silico* disease models at the very beginning of the drug discovery pipeline, as well as providing a high-performance simulation infrastructure for *in silico* testing of drug target candidates, generated from the hypothetical results of *in silico* disease models, will be of high value for overcoming such attritions.

5. Looking to the future

Direct extraction of biological information from text will certainly ease the curation process for databases, which is a challenging task for database annotators and domain experts. But it can be foreseen that in the not so distant future, textual data will be an indispensable part of "integrative biology" models, which aim at predicting biological outcomes by putting different components together.

The information encoded in the body of scientific literature has more to offer than can be found by traditional reading of publications one by one. The ability to look at hundreds of thousands of publications simultaneously and to do statistical analysis on factual statements in scientific text opens new perspectives for scientific work in the life sciences. For example,

another type of data that is of enormous potential in biomarker and target discovery corresponds to clinical outcome information, which reflects the physiological response to a drug or the diagnostic/prognostic value of specific biomarkers. So far such information in the text has been underutilized, although they offer complex descriptions of disease genotype and phenotype. Hence, there is a need to develop specialized terminologies for the extraction of clinical and biomarker information from the literature. Moreover, we expect that textual data embedded in the biomedical literature will play an important role in evidence-based approaches in medicine, such as empowering clinical decision-support systems by means of automated screening of scientific text for statements encoding medical evidence.

Notes

[i] http://www.eu-egee.org/

References

Banville, D.L. (2006), "Mining chemical structural information from the drug literature", *Drug discovery today*, Vol. 11 No. 1, pp. 35-42.

Botstein, D. and Risch, N. (2003), "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease", *Nature genetics*, Vol. 33 Suppl., pp. 228–237.

Butcher, E. C., Berg, E. L. and Kunkel, E. J. (2004), "Systems biology in drug discovery", *Nature biotechnology,* Vol. 22, pp. 1253-1259.

Chuang, H., Lee, E., Liu, Y., Lee, D. and Ideker, T. (2007) "Network-based classification of breast cancer metastasis", *Molecular systems biology*, Vol. 3:140, available at: http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=2063581&blobtype=pdf (accessed 16 April 2009).

Ergun, A., Lawrence, C. A., Kohanski, M. A., Brennen, T. A. and Collins, J. J. (2007), "A network biology approach to prostate cancer", *Molecular systems biology*, Vol. 3:82, available at: http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=1828752&blobtype=pdf (accessed 16 April 2009).

Foster, I. and Kesselman, C. (Ed.) (1999), *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann Publ., San Francisco, Calif.

Friedrich, C. M. *et al.* (2008), "@neuLink: A Service-oriented Application for Biomedical Knowledge Discovery", in Solomonides, T. (Ed.), *Global healthgrid*, IOS Press, Amsterdam, pp. 165-172.

Goh, K. I., Cusick, M.E., Valle, D., Childs, B., Vidal, M. and Barabasi, A. L. (2007), "The human disease network", *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 104 No. 21, pp. 8685-8690.

Hanisch, D., Fundel, K., Mevissen, H.-T., Zimmer, R. and Fluck, J. (2005), "ProMiner: Organism-specific protein name detection using approximate string matching", *BMC*

Bioinformatics, Vol. 6 Suppl.1, S14, available at: http://www.biomedcentral.com/1471-2105/6/S1/S14 (accessed 29 May 2009).

Hirschhorn, J. N. and Daly, M. J. (2005) "Genome-wide association studies for common diseases and complex traits", *Nature reviews. Genetics,* Vol. 6, pp. 95–108.

Hood, L., Heath, J. R., Phelps, M. E. and Lin, B. (2004), "Systems biology and new technologies enable predictive and preventive medicine", *Science*, Vol. 306**,** pp. 640-643.

Hopkins, A. L. (2008), "Network pharmacology: the next paradigm in drug discovery", *Nature chemical biology,* Vol. 4, pp. 682-690.

Ibison, P. *et al.* (1993), "Chemical literature data extraction: The CliDE project", *Journal of chemical information and computer sciences*, Vol. 33, pp. 338–344.

Iles, M. M. (2008), "What can genome-wide association studies tell us about the genetics of common disease?", *PLoS Genetics*, Vol 4 No. 2: e33, available at: http://www.plosgenetics.org/article/info:doi%2F10.1371%2Fjournal.pgen.0040033 (accessed 17 April 2009).

Jacq, N. *et al.* (2006), "Demonstration of in silico docking at a large scale on grid infrastructure", *Studies in health technology and informatics*, Vol. 120, pp. 155-157.

Jacq, N., Salzemann, J., Legré, Y., Reichstadt, M., Jacq, F., Medernach, E., Zimmermann, M., Maaß, A., Sridhar, V., Vinod-Kusam, K., Montagnat, J., Schwichtenberg, H., Hofmann, M. and Breton, V. (2008), "Grid enabled virtual screening against malaria", *Journal of grid computing,* Vol. 6 No. 1, pp. 29-43.

Jensen, L. J. *et al.* (2006), "Literature mining for the biologist: from information retrieval to biological discovery", *Nature reviews. Genetics,* Vol. 7, pp. 119–129.

Kasam, V., Zimmermann, M., Maaß, A., Schwichtenberg, H., Wolf, A., Jacq, N., Breton, V. and Hofmann, M. (2007), "Design of Plasmepsin Inhibitors: A Virtual High Throughput Screening Approach on the EGEE Grid", *Journal of chemical information and modeling*, Vol. 47 No. 5, pp. 1818-1828.

Kasam, V., Salzemann, J., Jacq, N., Mass, A. and Breton, V. (2007), "Large Scale Deployment of Molecular Docking Application on Computational Grid infrastructures for Combating Malaria", in Schulze, B. (Ed.), *Seventh IEEE International Symposium on Cluster Computing and the Grid: CCGrid 2007, 14-17 May 2007, Rio de Janeiro, Brazil,* IEEE Computer Society, Los Alamitos, Calif., pp. 691-700.

Kola, I. and Landis, J. (2004), "Can the pharmaceutical industry reduce attrition rates?", *Nature reviews. Drug discovery*, Vol. 3, pp. 711–716.

Konagaya, A. (2006), "Trends in life science grid: from computing grid to knowledge grid", *BMC Bioinformatics*, Vol. 7 Suppl 5, S10, available at: http://www.biomedcentral.com/content/pdf/1471-2105-7-S5-S10.pdf/ (accessed 8 April 2009).

Kotzin, S. (2005), "Journal Selection for Medline", paper presented at the 71th IFLA General Conference and Council: Libraries - A voyage of discovery, August 14th - 18th 2005, Oslo, Norway, available at: www.ifla.org/IV/ifla71/papers/174e-Kotzin.pdf (accessed 17 April 2009).

Krallinger, M., Erhardt, R. A. and Valencia, A. (2005), "Text-mining approaches in molecular biology and biomedicine", *Drug discovery today*, Vol. 10 No. 6, pp. 439–445.

Kruglyak, L. and Nickerson, D. A. (2001), “Variation is the spice of life”, *Nature genetics,* Vol. 27, pp. 234–236.

Lafferty, J., McCallum, A. and Pereira, F. (2001), “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”, in Brodley, C. E. (Ed.), *Machine Learning: Proceedings of 18th International Conference (ICML-2001)*, *Williams College, June 28 –July 1, 2001*, Kaufmann, San Francisco, Calif., pp. 282–289.

Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabo, G., Rual, J.-F., Fisk, C. J., Li, N., Smolyar, A., Hill, D. E., Barabasi, A.-L., Vidal, M. and Zoghbi, H. Y. (2006), “A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration”, *Cell,* Vol. 125, pp. 801–814.

Lu, X., Jain, V. V., Finn, P. W. and Perkins, D. L. (2007), “Hubs in biological interaction networks exhibit low changes in expression in experimental asthma”, *Molecular systems biology,* Vol. 3:98, available at: http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=1865580&blobtype=pdf (accessed 16 April 2009).

Mack, R. and Hehenberger, M. (2002), “Text-based knowledge discovery: search and mining of life-science documents”, *Drug discovery today,* Vol. 7, pp. 89-98.

McDaniel, J. R. and Balmuth, J.R. (1992), “Kekulé: OCR—Optical chemical (structure) recognition”, *Journal of chemical information and computer sciences,* Vol. 32, pp. 373–378.

Motulsky, A. G. (2006), “Genetics of complex diseases”, *Journal of Zhejiang University. Science B*, Vol. 7 No. 2, pp. 167-168.

Pujana, M. A., Han, J. D., Starita, L. M., Stevens, K. N., Tewari, M., Ahn, J. S., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B. *et al.* (2007), “Network modeling links breast cancer susceptibility and centrosome dysfunction”, *Nature genetics,* Vol. 39, pp. 1338–1349.

Rabiner, L. R. (1989), “A tutorial on hidden Markov models and selected applications in speech recognition”, *Proceedings of the IEEE,* Vol. 77 No. 2, pp. 257-286.

Rauwerda, H. *et al.* (2006), “The promise of a virtual lab in drug discovery”, *Drug discovery today*, Vol.11 No. 5-6, pp. 228–236.

Ruau, D., Kolarik, C., Mevissen, H.-T., Müller, E., Assent, I., Krieger, R., Seidl, T., Hofman-Apitius, M. and Zenke, M. (2008), “Public microarray repository semantic annotation with ontologies employing text mining and expression profile correlation”, *BMC Bioinformatics*, Vol. 9 Suppl. 10, O5, available at: http://www.biomedcentral.com/1471-2105/9/S10/O5 (accessed 29 May 2009).

Stevens, R., Glover, K., Greenhalgh, C., Jennings, C., Pearce, S., Li, P., Radenkovic, M. and Wipat, A. (2003), “Performing *in silico* experiments on the Grid: a users perspective”, in Cox, S. (Ed.), *Proceedings of UK e-Science All Hands Meeting, Nottingham 2-4 September 2003*, EPSRC, Swindon, pp. 43–50, available at: http://www.cs.ncl.ac.uk/publications/inproceedings/papers/682.pdf (accessed 20 April 2009).

Whittaker, P. A. (2004), “The role of bioinformatics in target validation”, *Drug discovery today: Technologies*, Vol.1 No. 2, pp. 125-133.

About the authors

Martin Hofman-Apitius is Head of the Department of Bioinformatics in the Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany. Martin Hofman-Apitius is the corresponding author and can be contacted at: martin-hofmann-apitius@scai.fraunhofer.de

Erfan Younesi is a Research Assistant in the Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany.

Vinod Kasam is a PhD student at the Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany.