Progress towards a comprehensive catalog of cancer genes based on gene-disease and gene-drug relationships identified

Kaj Albermann[1], Andreas Fritz[1], Karsten Wenger[1], Klaus Heumann[1], Peter A. Covitz[2], Lawrence W. Wright[2], Frank Hartel[2] and George A. Komatsoulis[2]

[1] Biomax Informatics AG, Lochhamer Str. 9, D-82152 Martinsried, Germany, [2] National Cancer Institute, Center for Bioinformatics (NCICB), 6116 Executive Blvd., Suite 403, Rockville, MD 20852, USA
Sophic Systems Alliance, Inc., One Research Court, Suite 450, Rockville, MD 20850, Program Director, Patrick M. Blake

An accurate and up-to-date inventory of cancer genes is a necessary foundation for advancing cancer research and supporting patient treatment. Clinical researchers and MDs treating patients all need easy access to a reliable core knowledge repository in order to connect "omics" with "oligies". The development and deployment of this data set requires a foundation built on common language and a standardized technology platform. The National Cancer Institute (NCI) is adressing these important issues by a) mapping biological terms in common use to unique concepts by building the NCI Thesaurus (http://ncicb.nci.nih.gov/download/index.jsp) and b) implementing the caBIG™ strategy to establish grid network standards in software interoperability to facilitate collaboration between cancer research centers. The emerging "touchstone" cancer gene index, combined with common data elements deployed to cancer labs and treatment centers, will accelerate the goal of trating and finding cures for cancer.

In 2003, NCI launched a cancer Gene Index Project to expand the cancer gene information in the NCI Thesaurus to its fullest extent. This project required the implementation of an automated and manual approach to systematically identify all possible human cancer genes found in Medline abstracts (Figure 1). The Medline® database[1] was analyzed for the co-occurence of cancer or drug terms with human gene names in single sentences. The automated analysisphase was performed using the Biomx BioLT Linguistics Tool and, after manual inspection, the initial findings identified 4,800 cancer genes in August 2004.

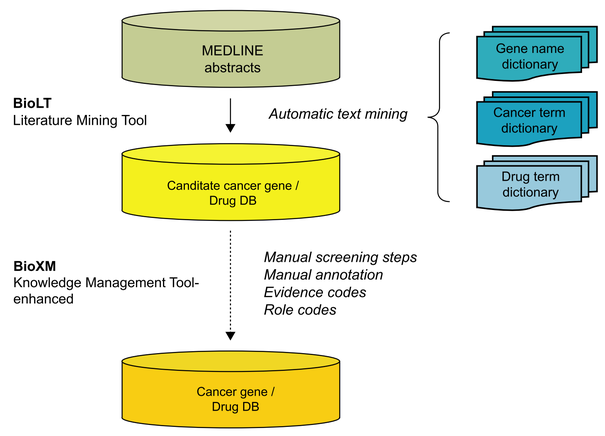

Figure 1. Project overview
Workflow for the extraction of all gene–cancer and gene–compound relationships from the current literature.

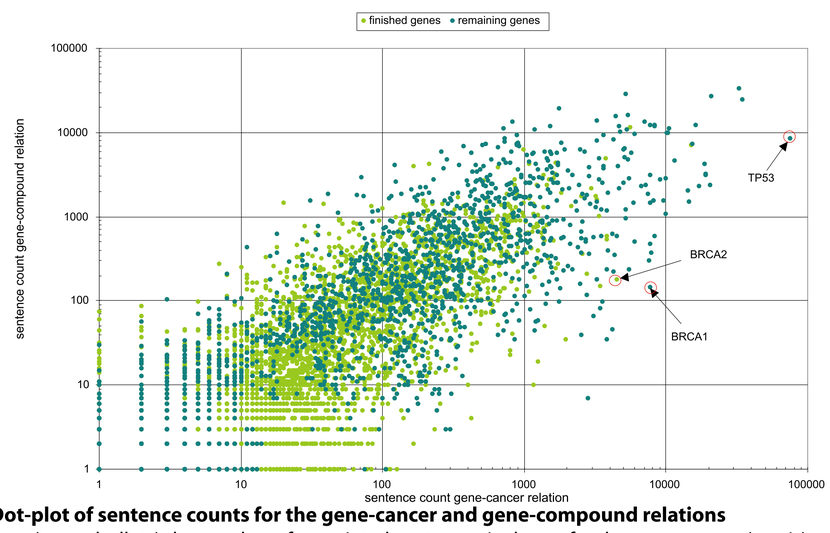

Figure 2. Dot-plot of sentence counts for the gene-cancer and gene-compound relations
For each gene (green bullets) the number of associated sentences is shown for the gene-cancer (x-axis) as well as the gene-compound (y-axis) relationship. Genes with cancer but no compound relationships are not displayed on this logarithmic graph. Cancer genes that already have been manually worked on are displayed in light green, genes that have not yet been worked on are shown in dark green. Three well known cancer associated genes are indicated. (TP53, tumor protein p53; BRCA1 breast cancer 1, early onset; BRCA2 breast cancer 2, early onset).

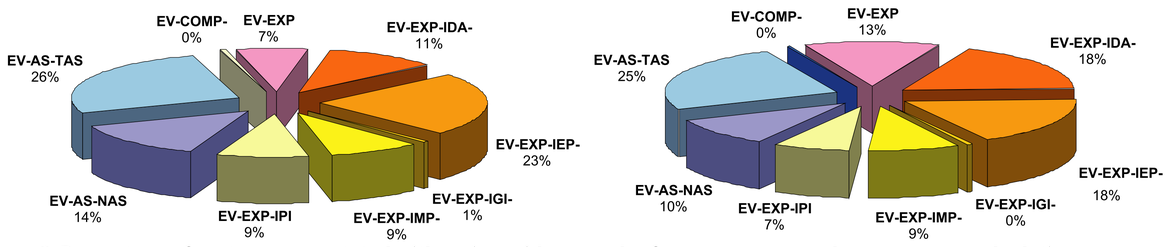

Figure 3. Percentage of sentences annotated with major evidence codes for gene-cancer and gene-compound relations
Left diagram: gene-cancer relations. Right diagram: gene-compound relations. More than one evidence code could be assigned for each sentence. In total 453,000 sentences containing gene-cancer relations were annotated. Of those about 231,000 sentences were classified as false positive (not shown in the graph). 556,000 sentences containing gene-compound relations were annotated. About 289,000 of these sentences were classified as false positive (not shown in the graph).
EV-AS-NAS: Non-traceable author statement; EV-AS-TAS: Traceable author statement; EV-COMP-: Inferred from computational analysis; EV-EXP: Inferred from experiment; EV-EXP-IDA-*: Inferred from direct assay; EV-EXP-IEP-*: Inferred from expression pattern; EV-EXP-IGI-*: Inferred from genetic interaction; EV-EXP-IMP-*: Inferred from mutant phenotype; EV-EXP-IPI: Inferred from physical interaction.*

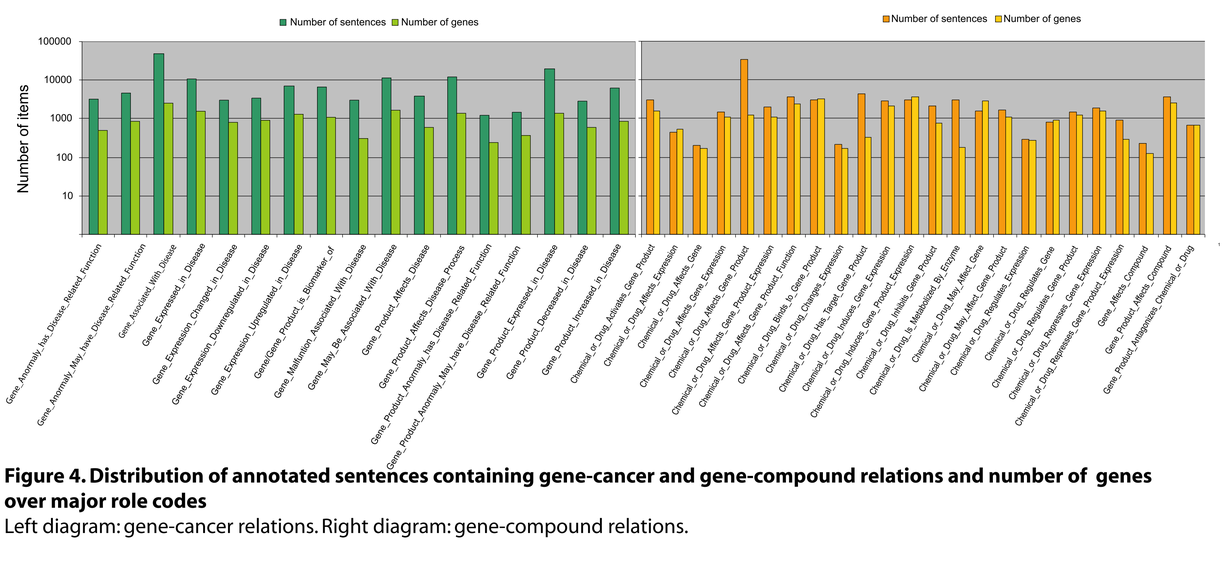

Figure 4. Distribution of annotated sentences containing gene-cancer and gene-compound relations and number of genes over major role codes
Left diagram: gene-cancer relations. Right diagram: gene-compound relations.

To date, we have automatically analyzed more than 12 million Medline abstracts from 1975 to 2005, revealing a total of some 10,000 distinct genes co-occuring with at least one cancer term. Through the ongoing analysis of the continuing publication of papers in Medline, the number of cancer genes has increased. after manual validation there are now 5,800 genes identified as having a relationship with cancer (Figure 2). Following the automated process and initial identification of genes, scientists perform manual annotation to identify, verify and validate relationships between the concepts 'gene', 'cancer types' and 'drug components' based on controlled vocabularies (Figures 3, 4). To date, 3,675 of the 5,800 identified cancer genes have been manually annotated, including assignement of evidence[2] and role codes[3] (Figures 5, 6), and delivers to NCI to be made available to the cancer research community.

This cancer gene data set derived from the Cancer Gene Index Project is now integrated with other high quality resources, such as the NCI Physician Data Query (PDQ) database[4] for clinical trials, in the first caBIG™ Bronze compliant commercial system: the Biomax BioXM Knowledge Management environment (Figure 7). This new convergence of high quality data-sets in BioXM and alignment with the caBIG Software Standards provides a broad and flexible foundation which allows research scientists and medical doctors to explore and better understand the gene/disease/compound/treatment relationships. Modelling, visualizing and manipulating these complex cancer gene networks will support research to find potential pathways and biomarkers.

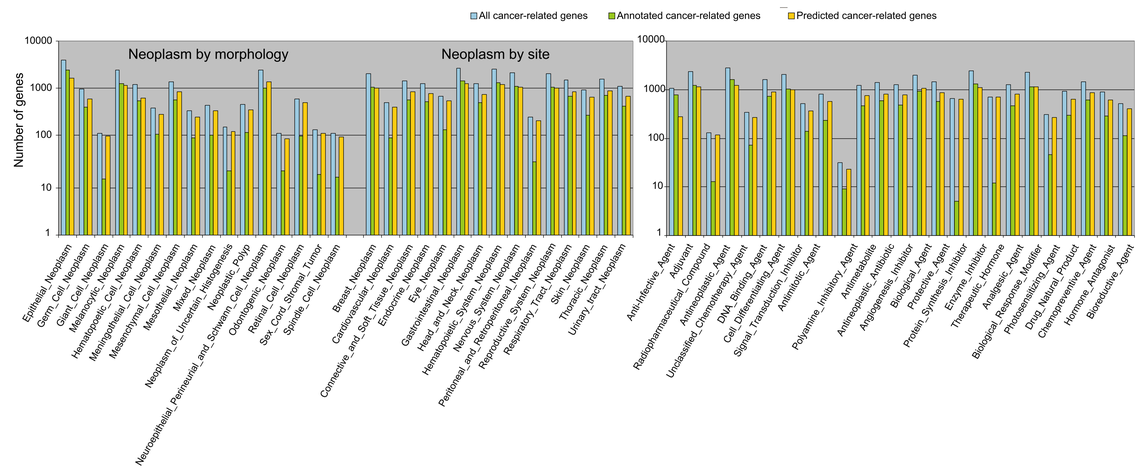

Figure 5. Distribution of genes over NCI thesaurus cancer and compound concepts
Left diagram: cancer concepts. Right diagram: Compound concepts.
For each concept the number of genes with at least one existing gene-cancer or gene-compound relation is shown for a) all genes that have been automatically classified or manually validated as being related to the respective cancer (all cancer-related genes), b) the genes that have so far been manually validated to have a relation to the respective cancer relation and have been annotated in detail (annotated cancer-related genes), and c) the genes that so far have not been worked on manually (predicted cancer-related genes).

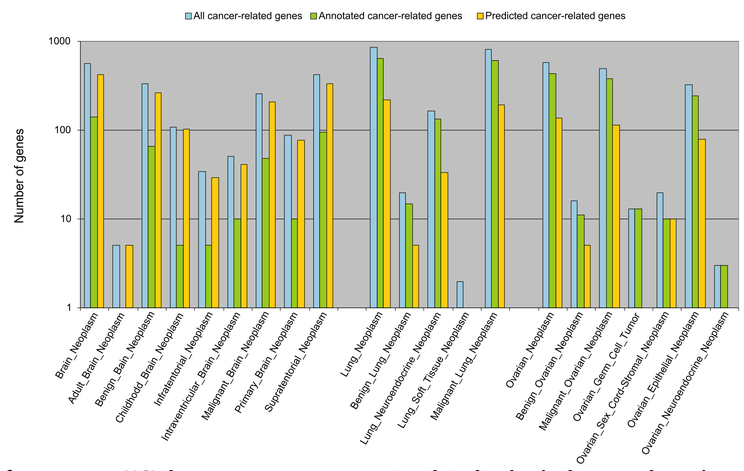

Figure 6. Distribution of genes over NCI thesaurus cancer concepts related to brain, lung, and ovarian cancer
For each concept the number of genes with at least one existing respective gene-cancer relation is shown for a) all genes that have been automatically classified or manually validated as being related to the respective cancer (all cancer-related genes), b) the genes that have so far been manually validated to have a relation to the respective cancer relation and have been annotated in detail (annotated cancer-related genes), and c) the genes that so far have not been worked on manually (predicted cancer-related genes).

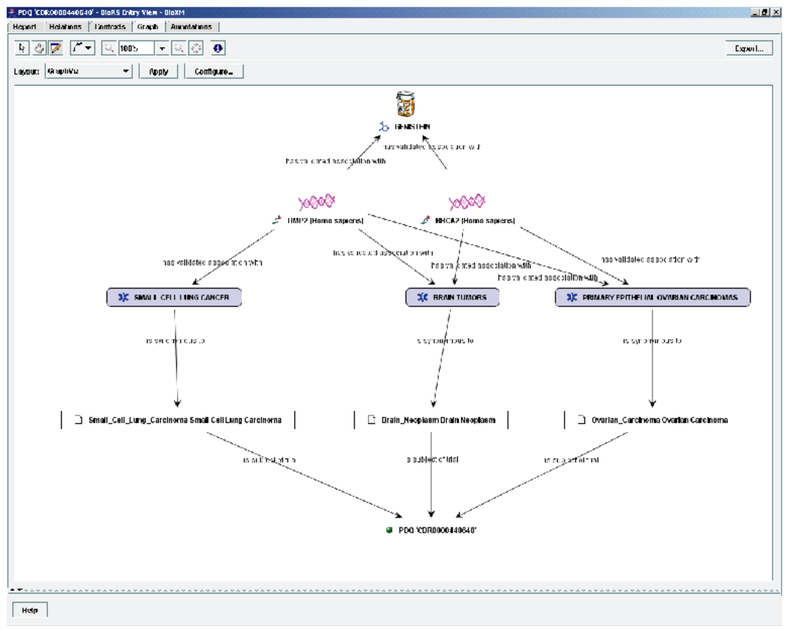

Figure 7. Integration of gene-cancer, gene-compound data with the NCI Physician Data Query (PDQ) database in the BioXM Knowledge Management Environment
As an example two genes (TIMP2, BRCA2) having relations to lung, brain, or ovarian cancer and compounds are shown. The screenshot depicts how relationships of genes, compounds, cancer, and clinical trials can be modelled as networked systems in the context of a clinically related project environment.

References
[1] *http://medline.cos.com/*
[2] *P.D. Karp, S. Paley, C.J. Krieger, and P. Zhang, PSB 2004 Online Proceedings. An Evidence Ontology for Use in Pathway/Genome Databases*
[3] *ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/March04current_roles.xls*
[4] *http://www.cancer.gov/cancertopics/pdq/cancerdatabase*

Contact
NCI Center For Bioinformatics
6116 Executive Blvd. Suite 403 - Rockville, MD 20850 - USA
+1 301 451-288
http://ncicb.nci.nih.gov

Sophic Systems Alliance Inc.
One Research Court Suite 450, Rockville, MD 20850 - USA
+1 301- 216 -3815
www.sophicalliance.com

Biomax Informatics AG
Lochhamer Str. 9, 82152 Martinsried - Germany
+49 89 895574-0
www.biomax.com