Towards a comprehensive catalog of gene-disease and gene-drug relationships in cancer

Christine M.E. Schueller[1], Andreas Fritz[1], Eduardo Torres Schumann[1], Karsten Wenger[1], Kaj Albermann[1], George A. Komatsoulis[2], Peter A. Covitz[2], Lawrence W. Wright[3] and Frank Hartel[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
[3] National Cancer Institute, Office of Communications, 6116 Executive Blvd., Rockville, MD 20852, USA

The National Cancer Institute (NCI) aimed to address the issue of mapping biological terms in common use to unique concepts by building their NCI Thesaurus[1], a reference terminology covering a broad area of basic and clinical science. This thesaurus contains nearly 110,000 terms in approximately 36,000 concepts partitioned into 20 sub-domains, which include diseases, drugs (compounds), anatomy, genes, gene products, techniques, and biological processes, among others, all with a cancer-centric focus in content. To expand the cancer gene section of the NCI Thesaurus to its full extent, a systematic approach was needed to collect all possible human cancer genes and to annotate the relationship of the concepts "gene", "cancer types" and "compounds" based on controlled vocabularies. NCI partnered with Biomax to enhance the cancer gene relationships contained in the NCI Thesaurus. As defined by NCI, the goal of the project was to produce a complete list of all cancer-related genes currently known in the public domain; defined by all validated relationships between the concepts "gene", "cancer" and "compound" found in the literature. Each relationship would include manual annotation that specifies the role of the gene in the corresponding cancer. It was necessary that these relations be extracted from the literature and annotated using controlled vocabularies (Table 1).

The project was organized into automatic and manual processes (Figure 1). The automatic process was designed to provide a superset of possible gene-to-disease or gene-to-compound relationships, while the subsequent manual steps ensured the specificity of the extracted relations. For automatic text mining, Biomax BioLT Literature Mining Tool was used to analyze the MEDLINE®[5] database for the meaningful co-occurrence of specific cancer or compound terms with human gene names. The results of the automatic text mining procedure were stored in a relational database and underwent manual validation and annotation. The goal of this step was to judge the automatic text mining results and to add annotation to the gene-cancer and the gene-compound relations.

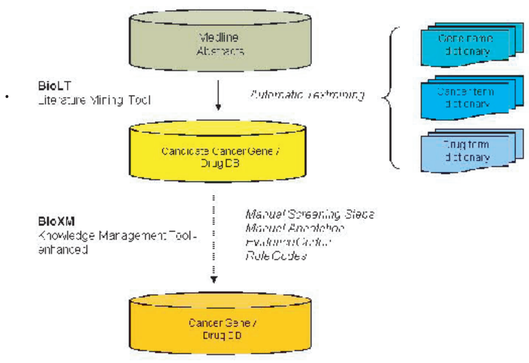

Figure 1. Project overview
Workflow for the extraction of all gene–cancer and gene–compound relationships from the current literature.

Vocabulary		Number of terms in vocabulary	Number of terms occuring in Medline (1975 to 06/2004)	Number of terms occuring in Medline (1975 to 06/2004)
Gene name	Synonyms	348,373	48,387	14%
	Unique	31,239	14,595	47%
Cancer term[2]	Synonyms	80,78	27,177	34%
	Unique	35,314	15,245	43%
Compound term[2]	Synonyms	11,817	2,79	24%
	Unique	3,254	1,651	51%
Evidence code[3]		35	-	-
Role code[4]		62	-	-

Table 1. Controlled vocabularies used in the project

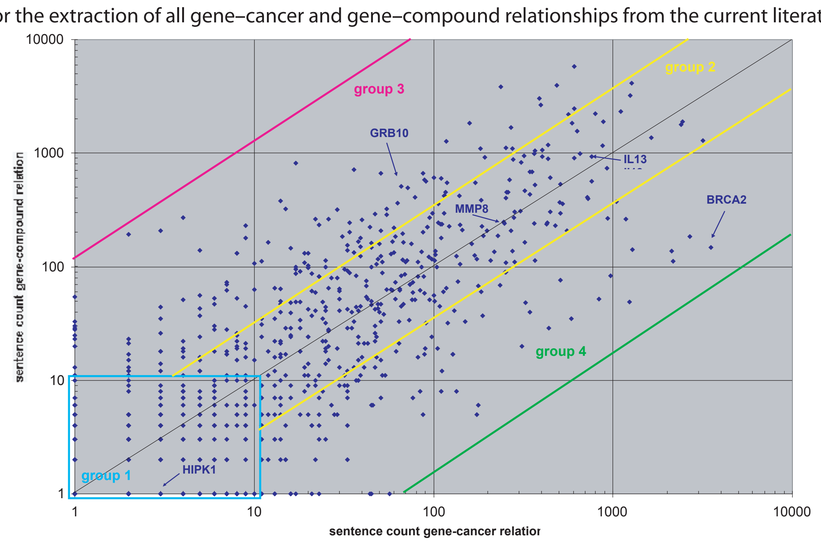

Figure 2. Dot-plot of sentence counts for the gene-cancer and gene-compound relations
For each gene (blue diamonds) the number of associated sentences is shown for the gene-cancer (x-axis) as well as the gene-compound (y-axis) relationship. Genes with cancer but no compound relationships are not displayed on this logarithmic graph. Cancer genes belong to one of four groups, depending on how many text relationships are found.
Group1: Few sentences (≤ 10) associated with both cancer and compounds.
Group2: Genes with extensive literature relating to both cancer and compounds.
Group3: Compound-focused genes.
Group 4: Cancer-focused genes.
One example gene belonging to each group is labelled (HIPK homeodomain interacting protein kinase 1; MMP8 matrix metalloproteinase 8; IL13 interleukin 13; GRB10 growth factor receptor-bound protein 10; BRCA2 breast cancer 2, early onset).

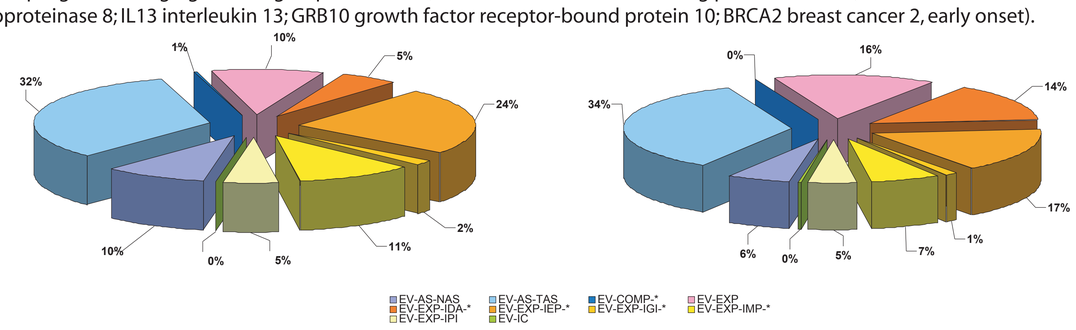

Figure 3. Percentage of sentences annotated with major evidence codes for gene-cancer and gene-compound relations.
Left diagram: gene-cancer relations. Right diagram: gene-compound relations. More than one evidence code could be assigned for each sentence. In total 89,598 sentences containing gene-cancer relations were annotated. Of those about 36,000 sentences were classified as false positive (not shown in the graph). 124,265 sentences containing gene-compound relations were annotated. About 69,000 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; EV-IC: Inferred by curator.*

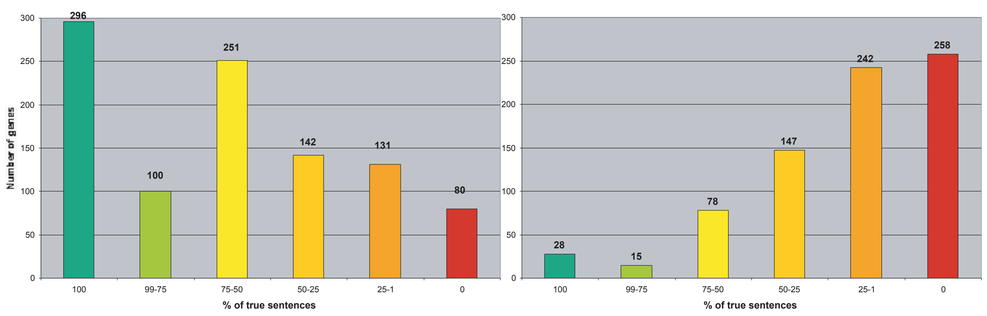

Figure 4. Correlation of number of genes and proportion of true sentences for gene-cancer relations and gene-compound relations
Left diagram: gene-cancer relations. Right diagram: gene-compound relations. The data were classified according to the proportion of true sentences per gene. Shown are the number of genes in each frequency class. True sentences have been assigned evidence codes other than 'false positive'. During the thorough annotation process another set of 80 genes was identified as being false positive cancer genes, as none of the associated sentences could be classified as a true statement.

The annotation service as well as the infrastructure for a robust, distributed work environment was provided by Biomax. The established framework also facilitated the use of various controlled vocabularies for manual annotation.
We screened 8.8 million MEDLINE abstracts, containing ~58 million sentences from 1975 to June of 2004 for occurrence of gene-cancer and gene-compound relations. In total about 17,000 gene names co-occurred with at least one cancer term, which reduces to approximately 8,000 individual genes. To identify genes that are truly associated to cancer, we manually screened the associated MEDLINE sentences for each gene. In a first rough screening, genes were scored as a true cancer related gene, when at least one sentence mentioning the gene together with a cancer term contained a true association. Of the 7,867 individual genes, 4,685 genes were scored as "true" cancer associated genes, 1,083 as "suspect" and 2,099 as "false". One thousand of the genes scored as being true cancer associated genes were selected randomly for a detailed manual annotation in a pilot study. The distribution of the number of sentences containing gene-cancer and gene-compound relations for the 1000 selected genes is shown in Figure 2.
The 1000 selected genes underwent a careful and thorough manual annotation process. For each of these genes, annotators read all the cancer-term and compound-term associated statements (sentences and/or abstracts) and added evidence and roles to each relevant statement. Evidence codes were assigned to qualify the assertion made in the statement in respect to the association of the cancer or compound term to gene name. Role codes describe, in general, the semantic association of the gene and the corresponding cancer or compound term (Figures 3, 4, 5, 6).

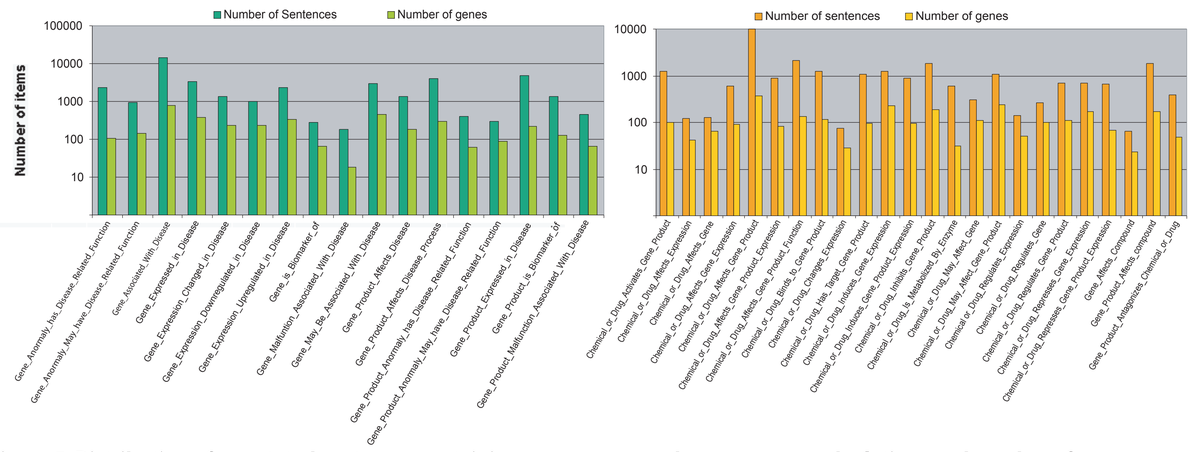

Figure 5. Distribution of annotated sentences containing gene-cancer and gene-compound relations and number of genes over major role codes
Left diagram: gene-cancer relations. Right diagram: gene-compound relations.

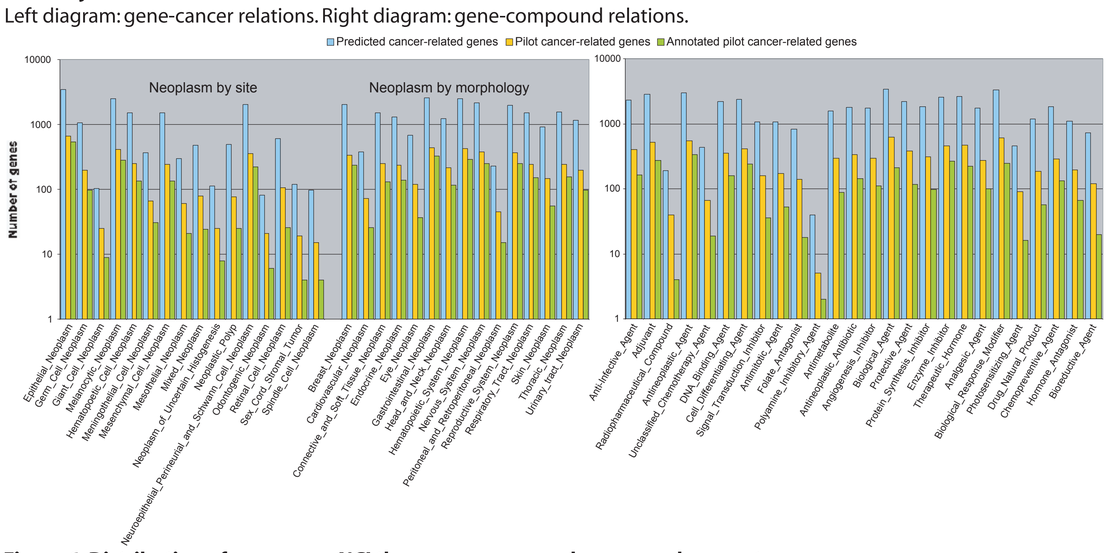

Figure 6. Distribution of genes over NCI thesaurus cancer and compound concepts
Left diagram: cancer concepts. Right diagram: Compound concepts.
For each concept the number of genes with at least one existing gene-cancer or gene-compound relation is shown for a) all 4,685 genes that have been classified as being true cancer genes by a first rough manual screening (predicted cancer-related genes), b) the 1000 genes in the pilot study before manual annotation (pilot cancer-related genes), and c) the 1000 genes in the pilot study after manual annotation (annotated pilot cancer-related genes).

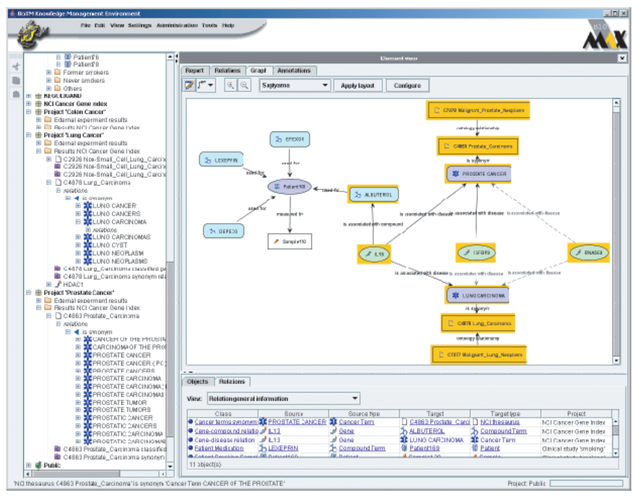

Figure 7. Integration of gene-cancer and gene-compound data in the BioXM Knowledge Management framework
As an example the IL13 gene and its relations to cancer types and compounds as well as other interesting entities is shown. The screenshot depicts how relationships of genes, cancer, and compounds can be modelled as networked biological systems in the context of a clinically related project environment.

Summary

As with many research groups, the NCI was faced with the problem of populating a reference terminology with biologically meaningful content so that interoperability and data sharing can facilitate company- or institute-wide projects. The NCI Thesaurus was improved by applying a strategy combining automatic text mining and intensive manual validation and annotation. Initially 8.8 million MEDLINE abstracts were extensively and thoroughly mined for significant associations of cancer terms, compounds, and genes. About 4,700 genes have been identified as being related to cancer. For 1000 genes all existing gene-cancer and gene-compound relations identified in MEDLINE sentences have been thoroughly validated and manually annotated.
The valuable cancer gene data set developed in this project will be available from the NCICB[6], but is also integrated in the BioXM system, an information and knowledge management system developed by Biomax (Figure 7). This allows scientists to model a disease area in relationship to current knowledge to obtain an integrated view of the processes in the context of networked biological systems, including metabolic pathways, signal transduction pathways and regulation. The combination of extensively annotated cancer gene data and a sophisticated knowledge management suite provides the scientists in the field with a unique research reference framework.

References
[1] *G. Fragoso, S. de Coronado, M. Haber, F. Hartel, and L. Wright, Comparative and Functional Genomics, 2005, in press*
[2] *http://nciterms.nci.nih.gov*
[3] *P.D. Karp, S. Paley, C.J. Krieger, and P. Zhang, PSB 2004 Online Proceedings. An Evidence Ontology for Use in Pathway/Genome Databases*
[4] *ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/March04current_roles.xls*
[5] *http://medline.cos.com/*
[6] *The NCI Center for Bioinformatics, http://ncicb.nci.nih.gov/*

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

Biomax Solutions Inc.
200 Main Street, Suite 201, Falmouth, MA 02536 - USA
+1 508 495-3801
www.biomaxsolutionsinc.com

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