First complete compendium of cancer genes
The Cancer Gene Index Project

Kaj Albermann[1], Andreas Fritz[1], Karsten Wenger[1], Klaus Heumann[1], George A. Komatsoulis[2], Juli D. Klemm[2], and Patrick M. Blake[3]

[1] Biomax Informatics AG ● Lochhamer Str. 9 ● 82152 Martinsried ● Germany
[2] NCI Center for Biomedical Informatics and Information Technology (CBIIT) ● 2115 E. Jefferson St., Suite 5000 ● Rockville, MD 20852
[3] Sophic Systems Alliance, Inc. ● One Research Court, Suite 450 ● Rockville, MD 20850

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".

In 2004, NCI launched the Cancer Gene Index Project (CGI) to provide the cancer community with a complete compendium of all cancer related genes occurring in the biomedical literature with manually annotated gene/disease and gene/compound relationships. The aim was to accelerate discovery of cancer drugs, biomarkers, cures and treatments.

A cancer gene was defined as any human gene or gene product that co-occurs in a single Medline® database[1] sentence with a cancer disease or compound/treatment term. The Biomax BioLT Linguistics Tool was used to automatically analyze the complete Medline database with more than 18M abstracts and 94M sentences.

During the 5-year project, 6,955 cancer genes were identified and 1.8M sentences were manually validated and annotated with role codes and evidence codes[2] for each gene/disease and gene/compound/treatment relationship by PhD scientists.

Sophic has integrated the Cancer Gene Index into the Biomax BioXM Knowledge Management Environment to support research at NCI's Center for Cancer Research. BioXM and Cancer Gene Index are configured to support translational medicine, biomarker, mRNA and pathway discovery use cases. BioXM was the first caBIG™ Bronze compliant commercial system and is configurable to Silver Compatibility.

The complete Cancer Gene Index data are available without restriction to the community through the National Cancer Institute's caBIO gene annotation database (https://wiki.nci.nih.gov/display/ICR/caBIO) or via bulk FTP download (https://cabig.nci.nih.gov/inventory/data-resources/cancer-gene-index) and Sophic's website (www.sophicalliance.com).

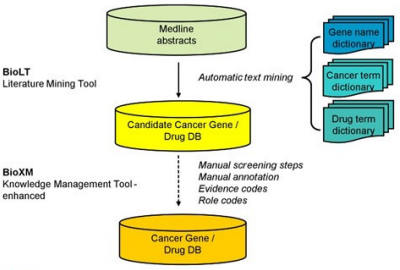

Figure 1. Project workflow.
Medline abstracts were automatically analysed for the co-occurrence of cancer and compound/drug terms with human gene names in single Medline sentences. For identified candidate genes , PhD. scientists performed manual annotations to identify, verify, and validate relationships between the concepts 'gene', 'cancer types', and 'drugs' based on controlled vocabularies.

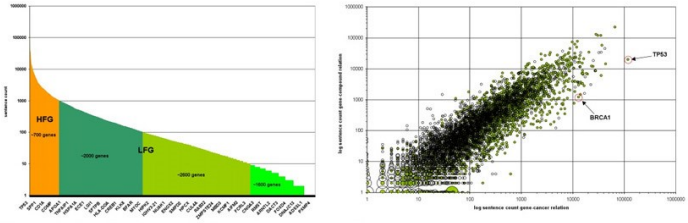

Figure 2. Classification of 6,955 cancer genes – Gene-cancer and gene-compound sentence counts
Left diagram - Well-studied cancer genes with over 1,000 Medline abstract sentences containing gene-disease relations in aggregate abstracts (e.g. TP53) were categorized as High Frequency Sentence Count Genes (HFG, orange) while less studied cancer genes with fewer than 1,000 sentences in aggregate abstracts were categorized as Low Frequency Sentence Count Genes (LFG, green). For LFGs, all sentences for "suspect" gene/disease and gene/compounds were manually curated. In contrast, the large number of sentences for HFGs required the development of a keyword-based prioritization approach to sentence curation. The approach ensured that all associations were captured without the need to recursively curate well-documented associations.
Right diagram - For each gene the number of associated gene-disease (x-axis) and gene-compound (y-axis) sentences are shown. Genes which are manually annotated to be potential biomarkers for cancer are displayed in green. The surface area of the circles is proportional to the number of genes. Two well known cancer associated genes are indicated (TP53, tumor protein p53; BRCA1 breast cancer 1, early onset).

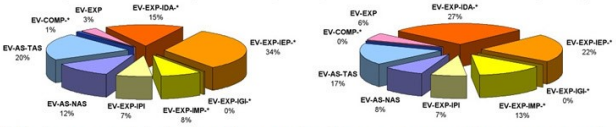

Figure 3. Sentences annotated with evidence codes for gene-cancer and gene-compound relations.
Left: gene-cancer relations. Right: gene-compound relations. Evidence codes qualify the origin of the assertion made in the sentence in respect to the association of a cancer or compound term to a gene. More than one evidence code could be assigned for each sentence. In total 977,000 sentences for gene-cancer relations were curated. Of those 535,000 were classified as false positive (not shown). 866,000 sentences containing gene-compound relations were annotated. About 547,000 of these sentences were classified as false positive (not shown).
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.*

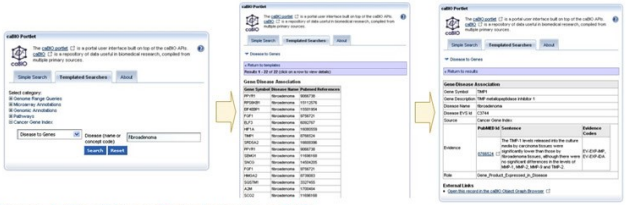

Figure 4. The Cancer Gene Index in caBIO.
caBIO is a repository of data useful in biomedical research compiled from multiple primary sources, including the Cancer Gene Index. The caBIO Portlet is a simple, user-friendly interface to caBIO and among the functions provided are queries for genes by an associated disease or agent and conversely, a query for diseases and agents associated with a given gene.
The caBIO Portlet is available at http://cagrid-portal.nci.nih.gov/web/quest/community .

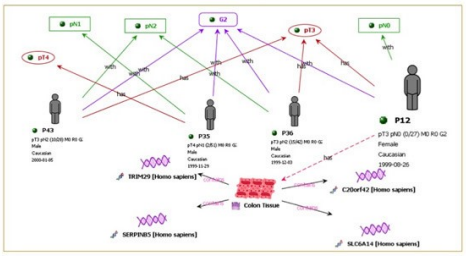

Figure 5. The Cancer Gene Index in BioXM - Translational medicine use case.
The focus of this use case is to support translational medicine research on colon cancer patients by integrating a full workflow of patient information from bed to bench. The system collects patient clinic histopathologic abnormality data (lymph node-pN, tumor size-pT and degree of metastases-G), tracks down patient tissue samples, conducts gene microarray analysis, uncovers unique expression patterns of colon cancer genes, and identifies literature supported evidences from the CGI project.

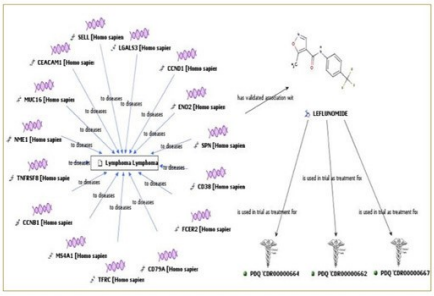

Figure 6. The Cancer Gene Index in BioXM - Biomarker use case.
BioXM knowledge relationship map for all CGI lymphoma cancer biomarker genes and lefunomide, a chemical compound related to the SPN lymphoma biomarker gene. The map is extended to show relationships between the lymphoma biomarker gene/leflunomide gene compound relationship and clinical trial information for lefunomide found in NCI's Physicians Data Query database. Researchers can explore complex disease, biomarker, compound and clinical trial relationships in a single GUI interface integrated with information in legacy systems, caBIG Software and information on the Grid.

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

Contact

NCI Center for Biomedical Informatics and Information Technology (CBIIT)
2115 E. Jefferson St., Suite 5000, Rockville, MD 20852 - 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