

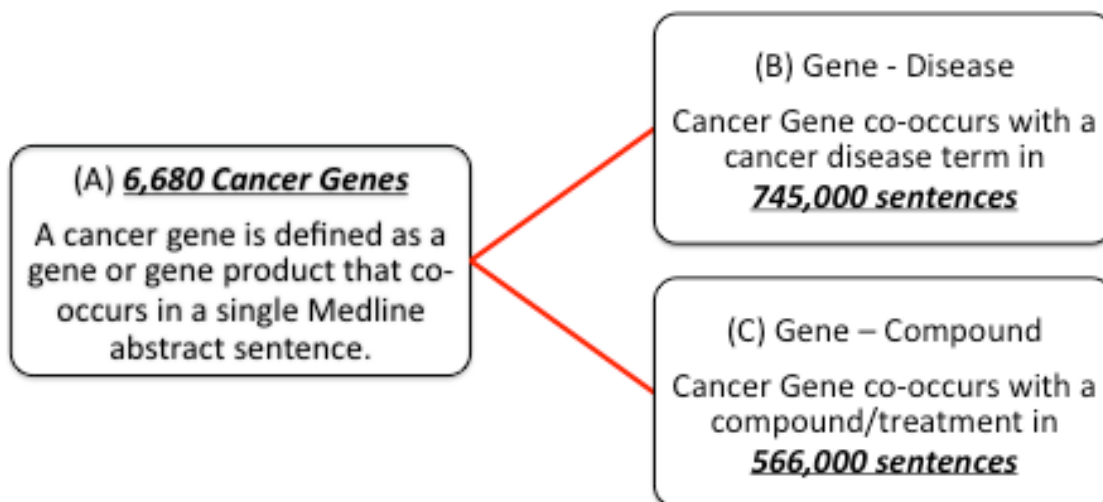
## White Paper NCI Cancer Gene Index Project

### Overview:

Since the first oncogene was discovered in the 1970s, scientists have aggressively pursued knowledge of cancer genes. As of October 2008, Medline contained over 18M abstracts with over 94M sentences related to biomedicine and a large number of these abstracts describe scientific results related to cancer. The objective has been to provide the cancer community with a complete index of all cancer related genes occurring in the biomedical literature in order to enable translational medicine, accelerate discovery of cancer drugs, biomarkers, cures and treatments. The Cancer Gene Index Project has produced the first single source of cancer gene information based on manually curated gene/disease and gene/compound relationships.

### Cancer Gene Index Data Model

*A "Bridge" to explore relationships between Disease to Biology to Chemistry*



- Approximately **7,000 new Medline abstracts** are published each day
- As of 10/10/08 Medline Refresh: **18M+ Abstracts, 94M+ Sentences**

The value of the CGI in accelerating the search for cancer treatments and cures is directly related to how much time and effort researcher's commit to learning to mine the index to find and understand relationships between scientific elements. Sophic has developed a suite of knowledge management configurations built on the CGI bridge data-model to support:

- translational medicine research
- biomarker discovery, the druggable cancer genome
- cancer gene abnormality and metastasis processes.

Examples of these use cases can be found on Sophic's website under Solutions ([www.sophicalliance.com/solutions](http://www.sophicalliance.com/solutions)).

Researchers should know the Cancer Gene Index is incomplete. Because of limited NCI funding, the project stretched over a 5-year period during which time, thousands of papers were published on newly discovered genes that are not included in the Cancer Gene Index. However, the index, as is, provides valuable, well organized, highly curated information on the majority of all cancer genes. The complete Cancer Gene Index will be funded, delivered and maintained based on the feedback from the cancer community to NCI on the value and use of CGI. It was a privilege for Sophic and Biomax to team with NCI on the development of the Cancer Gene Index.

### **Design:**

**In the Cancer Gene Index**, a cancer related gene is defined as any human gene or gene product that co-occurs in a single Medline abstract sentence with a cancer or compound/treatment term. A quality driven, "high throughput factory" method has been developed to combine automated natural language processing (NLP) and a team of highly trained Ph.D. scientists who perform manual annotation and curation of each cancer genes reviewed

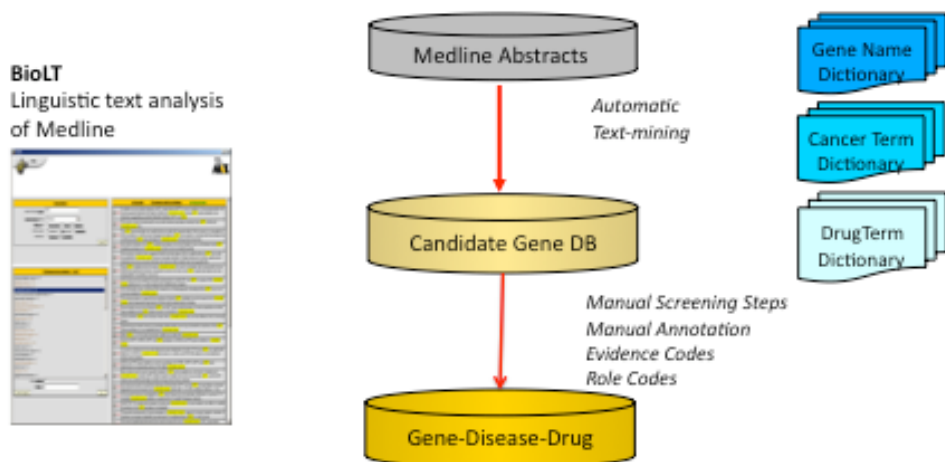
The automated NLP uses the National Cancer Institute (NCI) Thesaurus cancer and compound terms plus a human gene name dictionary to mine Medline abstracts for all "suspect" cancer genes that are then manually validated as "true" cancer related genes.

For each validated "true" cancer related gene, the cancer term and compound-term associated sentences are manually annotated with evidence codes and role codes from the NCI Thesaurus. Evidence codes qualify the origin of the assertion made in the statement in respect to the association of a cancer or compound term to a gene. Role codes describe the biological association between a gene and a cancer or compound term.

## Workflow Process:

### Cancer Gene Index Project Workflow Process

5-Year text-mining curation project: Mined 18M Medline abstracts, 94M sentences, identified 6,955 cancer genes, manually curated 1.3M sentences validating gene/disease & gene/compound/treatment relationships with NCI role codes & evidence codes.



- QA: Multiple quality control reviews throughout the “factory” annotation process.
- NCI QA: Reviewed and validated the quality of each set of annotated genes.

Sophic

Integrated Life Science Solutions

Biomax

### Solving the Problem of High Frequency Sentence Count Genes:

As of October 2008, Medline abstracts have over 4M sentences related to 310 High Frequency Sentence Count Genes. It is impractical to manually annotate all 310 High Frequency Sentence Count Genes (HFG). The logical design of a decision process for achieving the highest level of coverage with the limitations in time, resources and budget was complex. Below is the description of the methodology to provide useful, high quality sample annotations for the HFG. The methods and automated steps described below are available at Sophic to support ongoing updates of HFGs of specific interest being studied by researchers.

**Process (1):** Based on past project experience, we will use a NLP Drill down approach for selecting representative sentences from over 4M sentences describing 310 High Frequency Genes. (HFG means >1,000 sentences for a gene in PubMed Abstracts.)

**Decision (2):** Divide each of the HFG into 250,000 sentences for Gene-Disease and 250,000 Gene-Compound Relationship sentence relationship “batches”. Create a gene/disease and gene/compound relationship base line set of sentences to make sure we capture ALL gene-

disease and gene-compound relationships with at least 1 sentence. Estimated average of 3 sentences per relationships based on current analysis.

**Store Data (3):** Analysis 80% 310 HFGs. Gene-disease relationships: an initial statistical analysis was generated for the distribution of sentences over the NCI Thesaurus cancer codes for the 310 HFG. For each cancer code (e.g. C9385 Renal\_Cell\_Carcinoma) the numbers of sentences were counted that have been identified by NLP analysis to contain a gene-disease relation with the respective cancer code. Cancer codes were then grouped depending on the number of attached sentences. The resulting statistics showed that with a cut-off of 10 sentences attached to an individual cancer code about 80% (about 68,000 codes) of all cancer codes can be fully manually annotated. In total this represents about 185,000 sentences. In average this results in about 2.7 sentences per cancer code gene relation.

**Store Data (3):** Gene-compound relationships: the same approach was applied to analyze gene-compound relationships identified in sentences by NLP analysis. There a cut-off of 7 sentences attached to an individual compound code will allow about 75% (about 78,000 codes) of all cancer codes to be fully manually annotated. In total this represents about 188,000 sentences. In average this results in about 2.4 sentences per gene compound code gene relation to be annotated.

**Decision (4):** The manual annotation of the 185,000 Gene-Disease Relationships and 188,000 Gene-Compound Relationships can proceed.

**Process (5):** Manual annotation process of the 80% HGC Gene-Disease and Gene-Compound Relationships is work in process.

**Process (6):** Analysis of 20% Gene-Disease sentences of the 310 HFGs. **HOLD ON GENE-COMPOUND UNTIL GENE-DISEASE IS COMPLETE.** The remaining approximately 20% of the cancer and compound codes (about 16,000 and 28,000 codes) that are not covered by full manual annotation will be subject to additional NLP analysis in order to identify representative sentences for detailed manual annotation. For the gene-disease relationships there are about 65,000 sentences and for the gene-compound relationships about 62,000 sentences have to be selected.

To select these representative sentences, the NLP filters will require further development using a group-term method of identifying the most relevant sentences. For the Atlas project two logical categories with associated terms have been developed for the gene-disease relationships – abnormalities and expression. Two additional candidate categories are biomarkers and therapeutics. For the HFG gene-compound relationships candidate categories have to be developed.

The remaining highest sentence count GD – GC selection criteria and metrics for inclusion/exclusion will be discussed and agreed to over the next 4 weeks. NCI PM will provide the inclusion/exclusion criteria for selecting sentences to be manually annotated for the HFG genes - gene-disease relationships. The same criteria/metric will be used to determine inclusion/exclusion of the sentences representing gene-compound relationships. For the selection of sentences belonging to these categories a NLP approach will be applied. Dictionaries allowing selection of sentences belonging to these categories will be provided by Biomax PM and will be reviewed by NCICB Project Officer and COTR. The dictionaries should reflect recent shift in technology and focus in cancer research. Thus e.g. sentences containing information about epigenetic or SNP analysis should be identified.

**Decision (7):** Gene Disease Strategy – Explanation of HFG Gene Disease Sentence Selection based on the following statistical analysis:

- a. There are 84,739 gene-disease relationships for the 310 HF genes.
- b. 68,498 gene-disease relationships are covered by approximately 185,000 with less than 10 sentences for each distinct gene-disease relationships.
- c. There are 16,241 gene-disease relationships to be assigned to approximate 65,000 by a rational drill down approach.

**Decision (8):** For this approach 4 functional categories have been assigned:

- A Expression
- B Abnormality
- C Biomarker
- D Therapy

By NLP analysis 99% of the 16,241 gene-disease relationships are captured by sentences containing any of the 4 categories. In order to combine maximum value of sentence and broad coverage of the categories the sentences containing any combination of the categories has been determined by NLP analysis.

ABCD	AD
ABC	BC
BCD	BD
ACD	CD
AB	A
AC	B
	C
	D

The proposed approach based on the sentence count statistics is to move in three steps.

- 1) Annotate all sentences of the top 4 category combinations.
- 2) Pick the top 3-4 most recent publications for category combinations AB through CD for gene disease relationships not covered under 1)
- 3) Pick the top 3-4 most recent publications for category combinations A through D for gene disease relationships not covered under 1 or 2)

**Data (9):** Organize categories into 4 quadrants.

### 20% HFG Category & Drill-down Strategy

Q-1 ABCD – 341 Sentences	Q-2 ABC – 2,583 Sentences BCD – 574 Sentences ACD – 2,298 Sentences Q1 & Q2 Covers 12.4% of GD Rel.
Q-3 AB - 44,040 Sentences AC - 23,259 Sentences AD - 42,263 Sentences BC - 7,440 Sentences BD - 5,696 Sentences CD - 7,764 Sentences Another 61.5% GD Rel.	Q-4 A – 433,8099 Sentences B – 149,398 Sentences C – 44,655 Sentences D – 100,690 Sentences Another 25%

**Decision (10):** HFG – Gene-Disease Annotation Drill-down Strategy

**Process (11):** Brief the caBIG ICR on approach for selection sentences for Q3 and Q4 above.

**Decision 13:** Q-1 & Q-2 – Annotate all sentences.

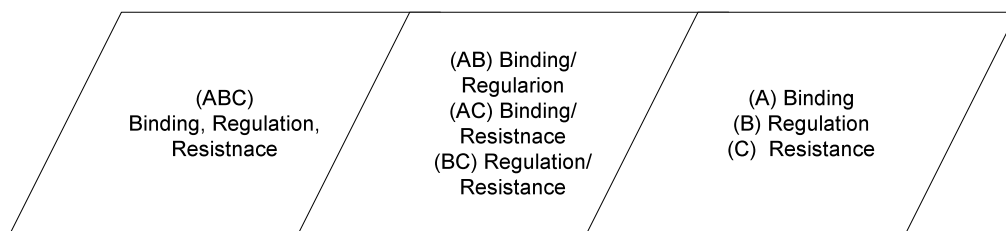
Q-3 & Q-4 – (A), (B) and (C) sequence decision process in Option 12.

**Process 14:** Proceed with Q1-Q2-Q3 and Q4 annotations of 20% HFG Gene-Disease relationships.

**Decision 15:** Proceed with using the 20% HFG Sentence Selection Methodology for Gene-Disease Relationships for 20% HFG Sentence Selection of Gene-Compound Sentences. Three categories of key terms have been selected for Gene-Compound Relationships.

- A. Binding
- B. Regulation
- C. Resistance

**Stored Data 16:** Key term categories for sentence relationships are grouped:



**Process: 17: HFG Gene-Compound Groups**

Group 1

ABC – Go forward on annotation.

Overlap sentences first

Use GD Decision Workflow to identify Group 2 & 3

Group 2

AB

AC

BC

Group 3

A

B

C

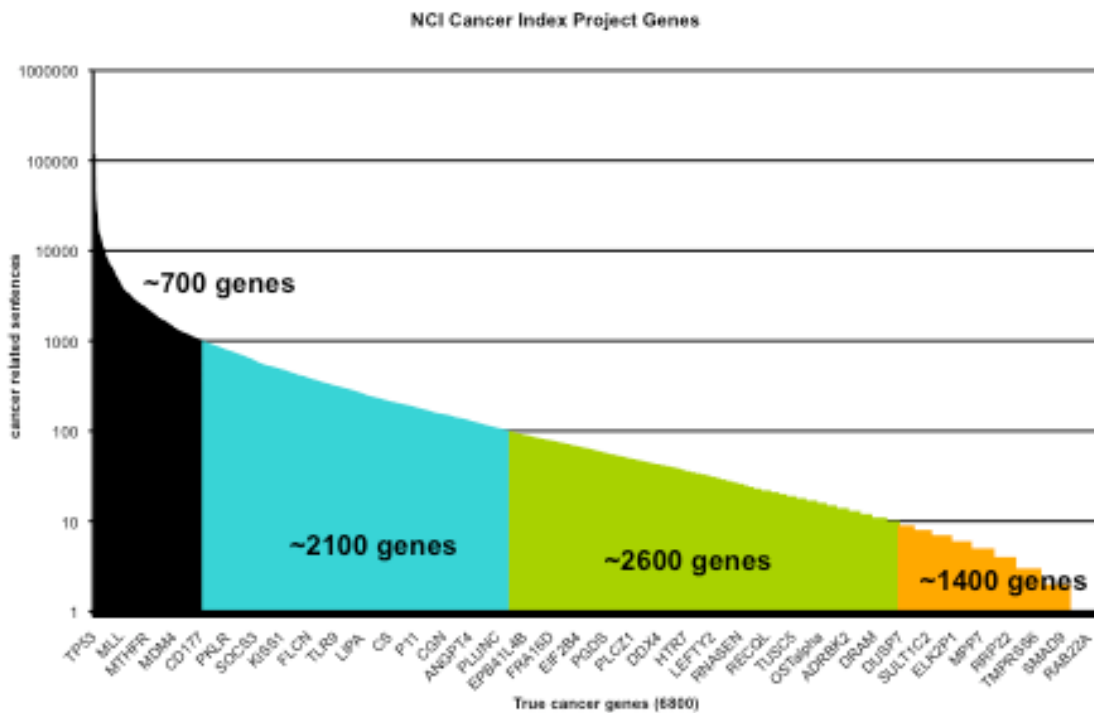
Group 1 first - ABC

Group 2 – 3 - Use decision sequence

**Results:**

As of April 2009, 6,955 published cancer genes have been validated as “true” cancer related genes, and about 2,200 of these genes were identified as biomarker cancer genes using the NCI Thesaurus Role Codes. Over 1.8M sentences were manually annotated to provide scientists a computable database, rich with detailed gene/disease, gene/compound cancer treatment and biomarker information on cancer genes.

**Distribution of the Gene-Sentence Statistics**



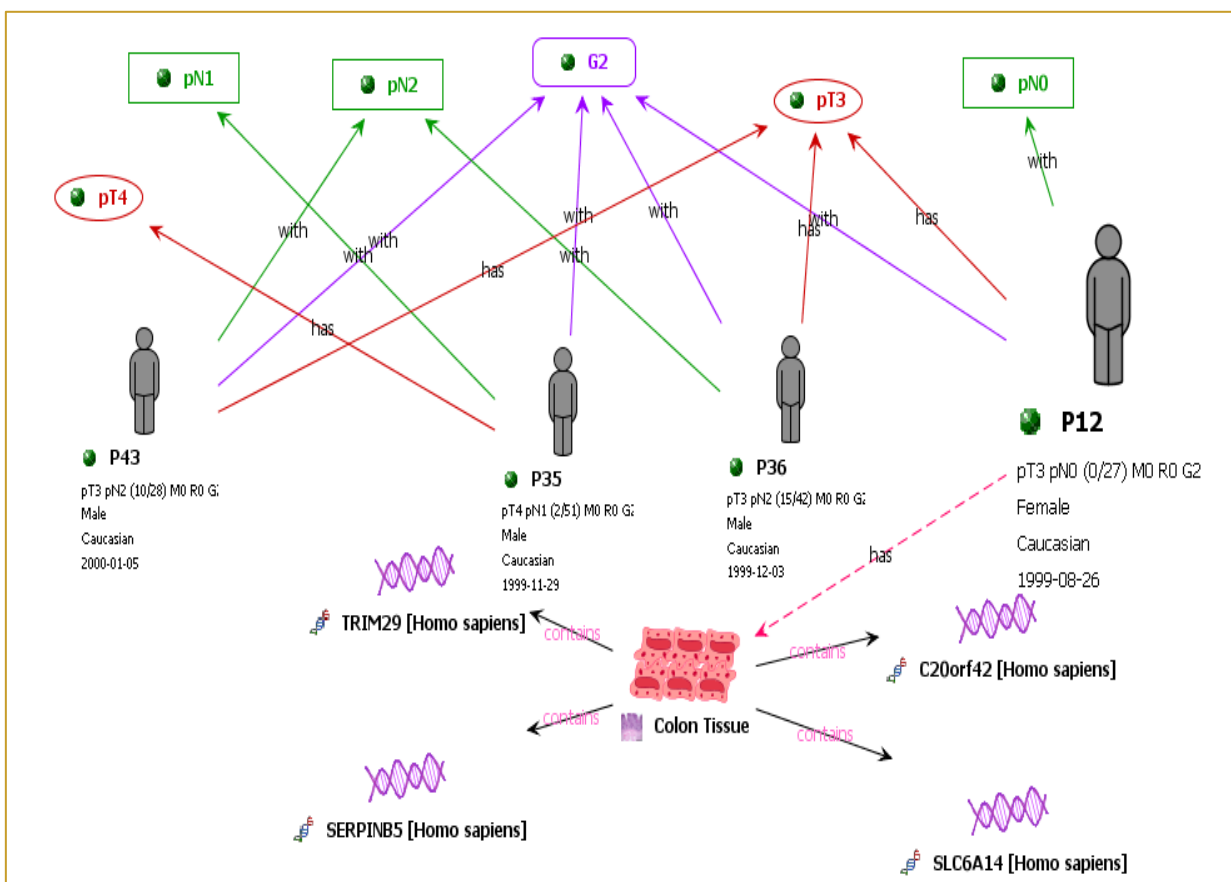
Sophic

Integrated Life Science Solutions

Biomax

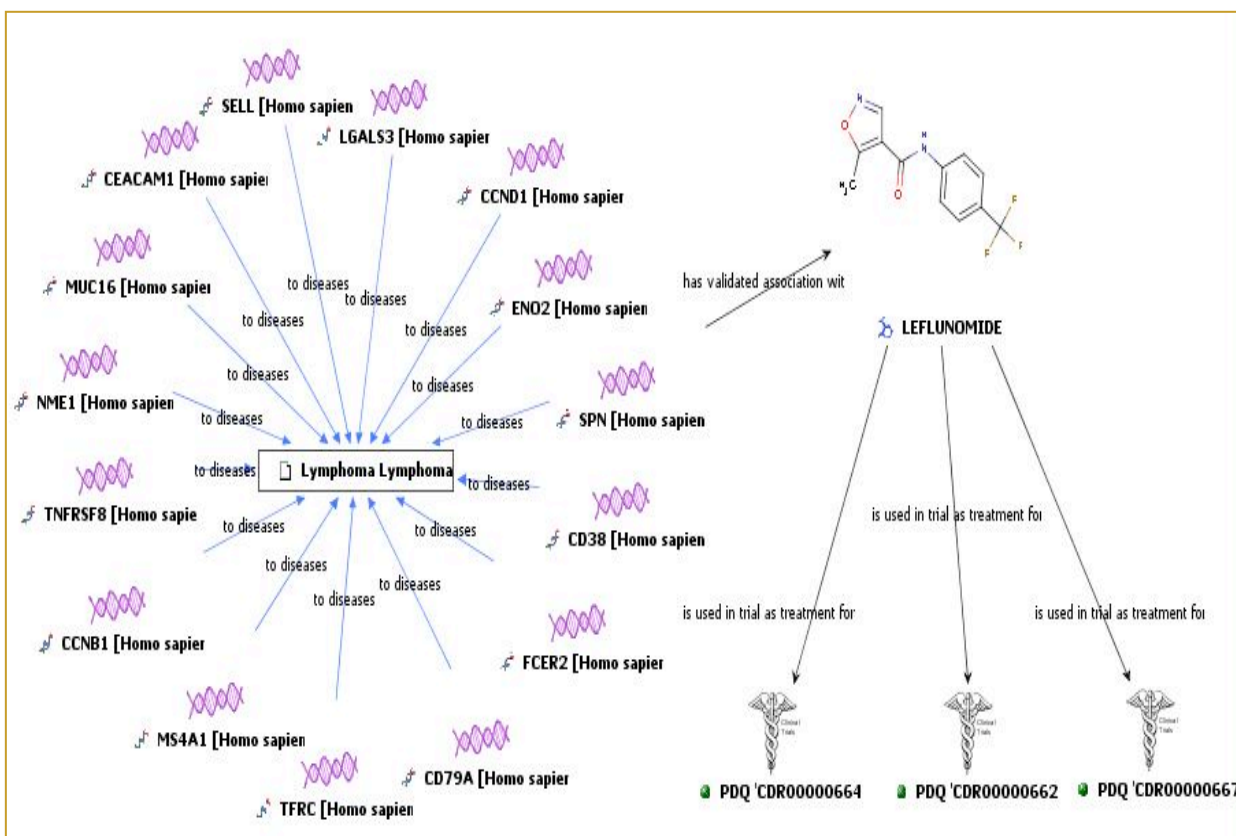
**Value of the Cancer Gene Index:**

The Cancer Gene Index provides the cancer community with the first semi-complete, single source of manually curated gene/disease and gene/compound association and will serve as a reliable integration foundation to support *in silico* cancer research. The use cases on the following pages are examples of how the Cancer Gene Index can be used as the data model backbone for translational medicine and biomarker discovery. Sophic configured the Biomax BioXM Knowledge Management System to create these use cases.



### The Cancer Gene Index Configured in BioXM Translational Use Case #1:

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.



## The Cancer Gene Index Configured in BioXM Translational Use Case #2:

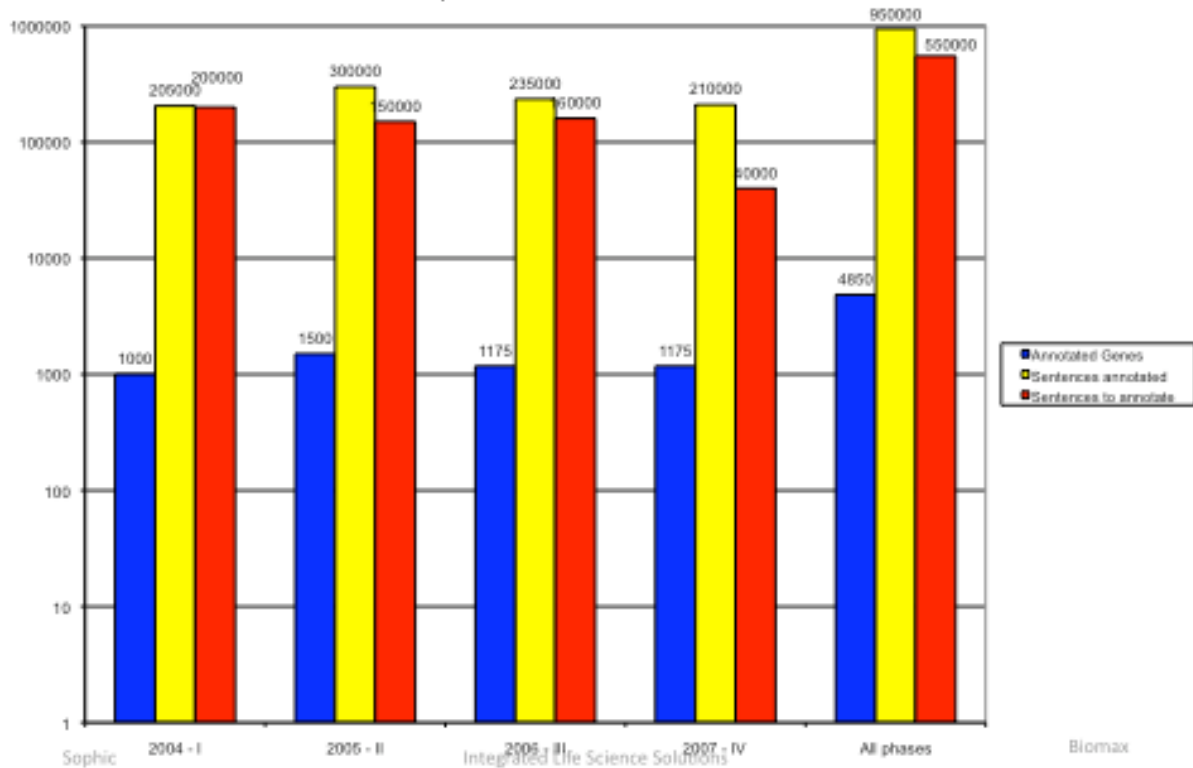
BioXM knowledge relationship map for all CGI lymphoma cancer biomarker genes and leflunomide, a chemical compound related to the SPN lymphoma biomarker gene. The map is extended to show relationships between the lymphoma biomarker gene/leflunomide gene compound relationship and clinical trial information for leflunomide 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.

**Current Status of the Cancer Gene Index:**

In 2009, approximately 7,000 new Medline abstracts were published each day, many related to new discoveries in cancer. With the surge of funding cancer research from the Stimulus, there will be a wave of new cancer genes and discoveries in 2010 and the years beyond. Because of the limited NCI funding for the Cancer Gene Index Project from 2005 – 2008, a gap was created in the annotation process for genes curated in Phases I, II, III and IV. The result is there are now more un-curated cancer papers on the early phase genes than there are newly published papers. This means that by now, over half of the cancer gene knowledge published is not included in the Cancer Gene Index. During the 5-year project, there were 4 updates of the corpus and linguistics analysis identified hundreds newly published cancer genes. The last corpus update was in January 2008 and it is logical to conclude that there is every day, significant numbers of “new true” cancer genes that are not included in the Cancer Gene Index.

**The Cancer Knowledge Wave**

Pre-publication - confidential information



Blue = Annotated Genes – Yellow = Annoted Sentences – Red = Sentences to be annotated

**The Cancer Gene Index Project Team:**

Kaj Albermann<sup>1</sup>, Andreas Fritz<sup>1</sup>, Karsten Wenger<sup>1</sup>, Klaus Heumann<sup>1</sup> George A. Komatsoulis<sup>2</sup>,  
Juli D. Klemm<sup>2</sup>, and Patrick M. Blake<sup>3</sup>

<sup>1</sup> Biomax Informatics AG • Lochhamer Str. 9 • 82152 Martinsried • Germany

<sup>2</sup> NCI Center for Biomedical Informatics and Information Technology (CBIIT) • 2115 E. Jefferson St., Suite 5000 • Rockville, MD 20852

<sup>3</sup> Sophic Systems Alliance, Inc. • One Research Court, Suite 450 • Rockville, MD 20850

**For more information on the Cancer Gene Index or the BioXM Knowledge Management Environment, contact:**

**Richard Navin**  
**Communications Associate**  
**Sophic Systems Alliance Inc.**  
**Email: [Richard@sophicalliance.com](mailto:Richard@sophicalliance.com)**  
**Office: 508-495-3801**  
**Cell: 508-564-0437**