Biomax Cancer Gene Database

Built in partnership with NCI

1. Introduction

Every scientist involved in basic cancer research, target-gene identification, target validation, chemical compound development, and cancer diagnostics and treatment has probably already spent days, weeks or even months mining the current scientific literature for information which correlates genes with diseases and drug compounds. The National Cancer Institute (NCI) saw this problem as a major obstacle to research progress and selected Biomax to produce the NCI Cancer Gene Database of all human cancer-related genes to be integrated in the caCORE[1] infrastructure for cancer informatics. Building on the current NCI Cancer Thesaurus,[2] we used the powerful Biomax BioLT™ Linguistics Analysis Tool to mine MEDLINE[3] for all known cancer-related genes and drug compounds used to interfere with the deadly effect of the disease. Manual annotations of the gene–disease and the gene–drug relationships were performed with an enhanced infrastructure based on the Biomax BioXM™ Knowledge Management Environment. In the first project phase, all gene–disease and gene–compound relations for 1,000 cancer-related genes have been manually validated and annotated. The project continues, and completion of this curation process for all cancer genes is planned for next year.

2. The problem

Today, there is no single complete data source containing all known cancer-related genes. Further, there is no process for updating the newly emerging cancer information continuously published in MEDLINE. While many lists of cancer genes are publicly available or reside within proprietary databases, such lists require diligent maintenance and constant updating. For example, one of the widely used cancer gene databases at Infobiogen[4] in France currently contains 2,648 entries (as of December 2004), but the question remains whether the information is complete and current. Each day, an average of 3,000 additional papers are published, many of which contain information on cancer. An examination using linguistic analysis of the complete MEDLINE database, consisting of approximately 11 million abstracts in December 2004, revealed that the scientific literature mentions over 8,000 genes in conjunction with cancer. Are these additional 5,500 genes simply naming or spelling variations, outdated gene names or cancer relations falsely recognized by the linguistic analysis? Or are some of these genes the hidden treasure that scientists studied at some point in time, but for which follow-up studies never occurred?

Indeed, one problem which makes it difficult for a curator to keep gene repositories consistent and up-to-date is the creativity and flexibility scientists have to constantly invent new names for genes and diseases. Efforts to standardize gene names have led to the replacement of old names already in use, which did not follow the conventions. However, many scientists still use the old, familiar names in publications. As a consequence, specialized gene name databases must keep track of all aliases for a name of a certain gene.

3. NCI Thesaurus

The NCI aimed to address the issue of mapping biological terms in common use to unique concepts by building the NCI Thesaurus.[5] This reference includes terminology covering cancer-specific genes as well as the much broader areas of basic and clinical sciences. The thesaurus contains nearly 110,000 terms in approximately 36,000 concepts partitioned into 20 sub-domains. The sub-domains have a cancer-centric focus in content and include diseases, drugs, anatomy, genes, gene products, techniques, and biological processes, among others. Each concept represents a unit of meaning and contains annotation, such as synonyms and preferred name as well as textual definitions and optional references to external authorities. In addition, concepts are modeled with description logic (DL) and defined by their relationships to other concepts.

With far fewer than 8,000 cancer-specific entries in the concept "gene", this section of the NCI Thesaurus clearly covers only a portion of all cancer genes. To expand the thesaurus to its full extent, a systematic approach was needed to collect all possible cancer genes and to annotate the relationship of the concepts "gene," "cancer types" and "drug components" based on controlled vocabularies used in the NCI Thesaurus.

[1] Covitz PA et al (2003) caCORE: A common infrastructure for cancer informatics. *Bioinformatics*. 19:2404-12
[2] http://nciterms.nci.nih.gov
[3] http://medline.cos.com/
[4] http://www.infobiogen.fr/services/chromcancer/Genes/Geneliste.html
[5] Fragoso G et al (2004) Overview and Utilization of the NCI Thesaurus. *Comp Funct Genom*. 5: 648–654

4. Project design and background

NCI recognized the above shortcomings as a major obstacle to research progress and partnered with Biomax to complete the cancer gene section of their 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. This list includes all validated relations between the concepts "gene," "cancer" and "drug" found in the literature as well as manual annotation that specify the role of the gene in the corresponding type of cancer. It was necessary that these relations be extracted from the literature and annotated using controlled vocabulary.

The project was organized into automatic and manual processes. This two-step approach allowed the automatic text-mining step be performed with high sensitivity to avoid omitting any possible relation. At the same time, the subsequent manual step ensured necessary specificity of the extracted relations.

For automatic text mining, the Biomax BioLT Linguistic Analysis Tool was used to analyze the MEDLINE database for meaningful co-occurrences of specific cancer or drug terms with human gene names.

The BioLT application is a linguistic text-mining tool that analyzes free text and extracts relations between various fields of interest (e.g., genes and diseases) in a global manner. It uses a classical lexical approach to identify specific terms in the text and returns related sentences and abstracts in a structured, easy-to-read display.

The terms that the BioLT application looks for are taken from dictionaries, which are first built from the text corpus under analysis. This procedure ensures a high precision and recall of term detection in free text. For example, the cancer term dictionary incorporates a number of public domain catalogs and classifications. However, it is necessary that the terms in the dictionary actually match the terms used in the literature and vice versa. Therefore, catalogs are augmented with a corpus-based extraction method and other procedures such as identification of morphologic variations, acronym recognition and disambiguation. Because the cancer term dictionary is a database for the purpose of term detection in free text, but not a well designed ontology (such as the NCI Thesaurus), a combination of both knowledge bases was a very attractive approach. After mapping and comparing the cancer-term section of the NCI Thesaurus with the BioLT cancer-term dictionary, both knowledge bases were combined in a non-redundant manner, such that all terms from both sources that are represented in the literature were retained.
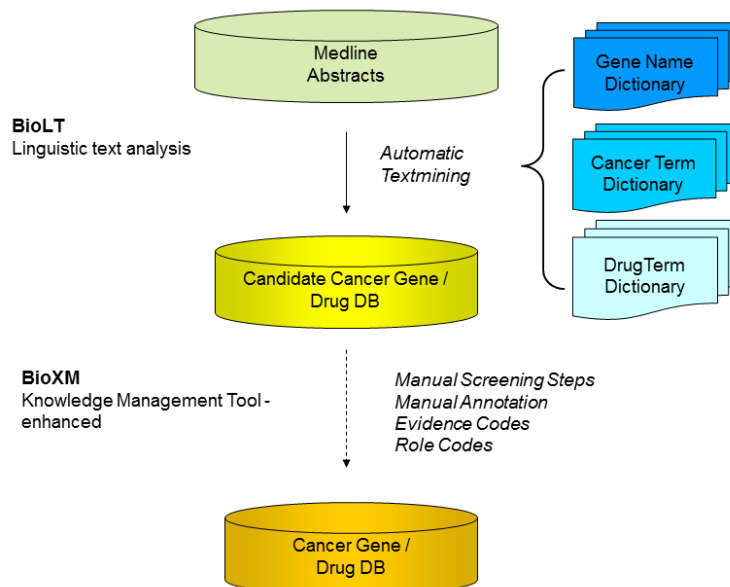

Figure 1. Workflow for the extraction of all gene–cancer and gene–drug relationships from the current literature

The BioLT gene name dictionary was established based on a unified database consisting of LocusLink[6], HGNC[7], and GDB[8]. Additional discovery procedures were used to broaden this initial database, ensuring a better mapping of dictionaries and literature. These sophisticated procedures include identification of morphologic

---

[6] http://www.ncbi.nlm.nih.gov/LocusLink/
[7] http://www.gene.ucl.ac.uk/nomenclature/
[8] http://www.gdb.org/

variations, acronym recognition and disambiguation and context-based gene name recognition. For the drug term dictionary, terms were taken from the NCI Thesaurus.

The results of the automatic text-mining procedure were stored in a relational database and underwent manual validation and annotation. Goals of this step were to judge the automatic text-mining results and to add annotation to the gene–cancer and the gene–drug relations.

As infrastructure for this second step in the construction of this cancer gene index, an enhanced and extended version of the Biomax BioXM Annotation Module was employed. The module is part of an integrated knowledge management suite and was built for standard annotation processes in research labs. This extended version was adjusted to handle the massive and continuous annotation process required to produce a database the size and complexity of the Biomax Cancer Gene Database. It provided the necessary framework for a distributed annotation environment robust enough to support a large number of annotators and also facilitated the use of various controlled vocabularies for the manual annotation.
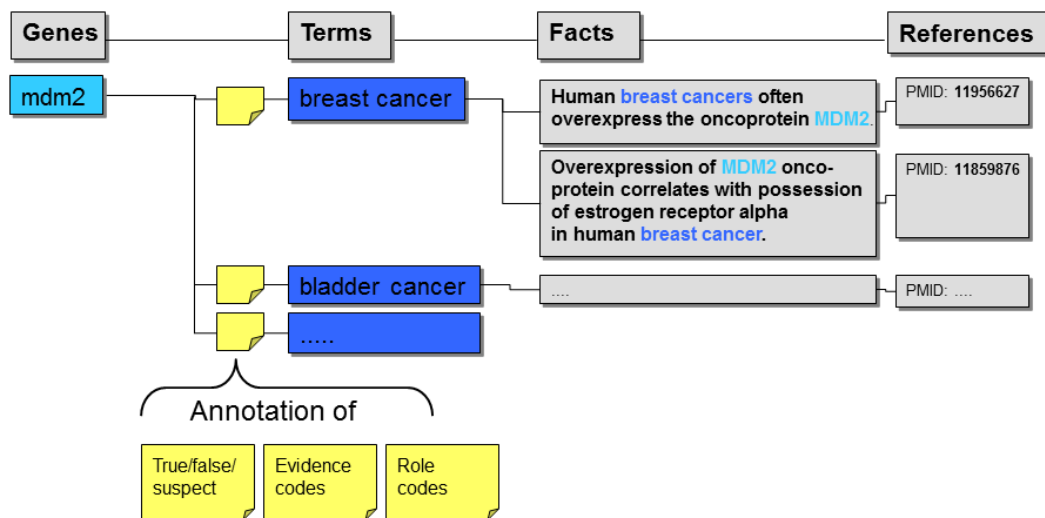

Figure 2. Manual validation and annotation of the gene–disease relations

An initial manual screening removed false positive cancer-related gene matches. This screening was based on the display of the related cancer terms for each gene as well as the attached sentences and abstracts that indicate a relation. Each true gene–cancer relation was then tagged accordingly.

A prerequisite of this project was to apply consistent terminology as used in the current NCI Thesaurus. Hence, only controlled vocabularies were used for the annotation of both the gene–cancer as well as the gene–drug relationships. One such vocabulary was the evidence codes,[9] which describes the scientific evidence of the captured information. Another was the NCI role codes,[10] which describe the semantic associations among the entities "gene," "cancer" and "drug." These entities are called "concepts" in the NCI Thesaurus. "Roles" are binary associations between concepts pairs.

5. Solution to the problem

This thorough and consistent annotation described above provides NCI with a valuable data source to populate the NCI Thesaurus.

In detail, the linguistic analysis of approximately 8.8 million abstracts of the MEDLINE database published from 1975, when the first oncogene was discovered, to 2004, when the study was performed, revealed a total of some 8,000 potential cancer genes. Input for this analysis consisted of 350,000 entries in the human gene name dictionary and 80,000 entries in the cancer term dictionary. Of these, 4,800,000 occurrences of cancer terms and 12,600,000 occurrences of gene names were found in the texts of these MEDLINE abstracts. In addition, 17,000 gene names co-occurred with at least one cancer term. These gene names reduce to approximately 8,000 distinct genes. After manual inspection, 4,800 of these were identified as true cancer-related genes. This number is nearly twice what other databases[3] list, although it may change slightly due to consistency checks as we continue the project.

---

[9] Karp PD, Paley S, Krieger CJ and Zhang P (2004) An Evidence Ontology for Use in Pathway/Genome Databases. *PSB*. 2004 online proceedings
[10] ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/March04current_roles.xls

The distribution of the number of sentences attached to each of these genes is shown in the figure below. The graph indicates a wide range of sentence counts per gene: from a single sentence per gene to more than 40,000 sentences for the p53 gene. Five genes have more than 10,000 sentences attached, and approximately 200 genes are mentioned in more than 1,000 sentences. Approximately half of all genes are mentioned in only ten or fewer sentences.

For the first project phase 1,000 genes with true cancer relation were annotated manually. These were derived from different regions of the graph in order to maximize the return on this project. Furthermore, it served as an insightful exercise to select genes based on their initial year of publication in order to obtain new and interesting insights as well as to capture important observations from the early years of cancer research.
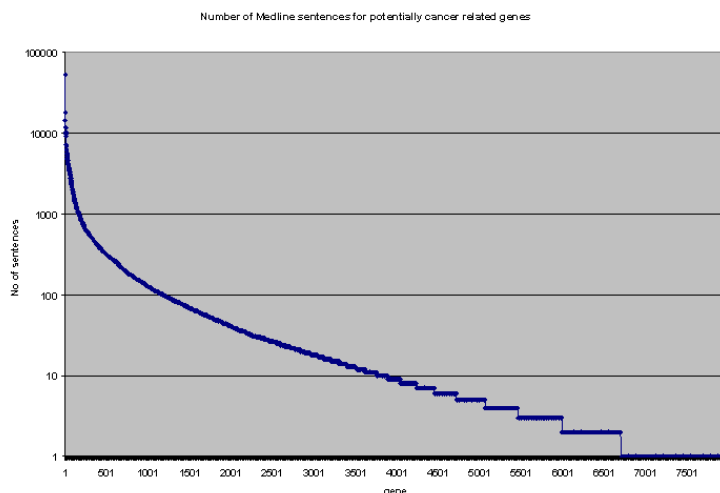
Number of Medline sentences for potentially cancer related genes

Figure 3. Distribution of sentence count per gene

For each of these genes, annotators read all the cancer-term-associated and drug-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 sentence with respect to the association of cancer or drug term to gene name. Role codes describe, in general, the semantic association of the gene and the corresponding cancer or drug term. The role codes were annotated in a way that the following combination links concepts and produces a meaningful sentence: gene name ↔ role code ↔ cancer or drug term.

It is obvious that these detailed annotations could be performed only by reading through every single sentence and abstract that was attached to a certain gene–cancer or gene–drug relation. In total, nearly 85,000 sentences had to be screened manually for the gene–cancer relation and about 120,000 sentences for the gene–drug relation.

With the completion of the first project phase, the results for the first 1,000 genes matched to specific diseases and chemical compounds is scheduled to be posted on the NCI website[11] in upcoming releases of caCORE. The project continues, and completion of this curation process for all cancer genes is planned for next year.

6. Conclusion

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 with the successful application of the BioLT linguistic analysis, which extensively and thoroughly mined the free text from 8.8 million MEDLINE abstracts for significant associations of cancer terms and gene names. The unique abilities of Biomax to take an initial framework of the customer and fill that in with both thorough (automatic) text mining and consistent and careful manual annotation of those text associations provided NCI a solution in a timely manner. The partnership of expertise that made this possible cannot be undervalued when a complete cancer gene database is produced. The integration of Biomax' abilities and the customer's vision can be applied to many other challenges in biological knowledge management in complex diseases as well as other areas of interest.

Moreover, the valuable cancer gene database that was developed in the course of this project will be integrated in the BioXM Knowledge Management Environment, an information and knowledge management system,

---

[11] http://ncicb.nci.nih.gov/NCICB/core

developed by Biomax during the last year. The underlying idea of the BioXM Knowledge Management Environment is to allow 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, on the one hand, and a sophisticated knowledge management suite like this, on the other hand, provides the scientists in the field with a unique research reference framework.

7. Contact Biomax

For more information about the Biomax Cancer Gene Database, contact Biomax Informatics AG.

Biomax Informatics AG
Robert-Koch-Str. 2
D-82152 Planegg
Germany
Tel.: +49 89 895574-0
Fax: +49 89 895574-825
Email: info@biomax.com
Website: www.biomax.com