DRUG DISCOVERY NEWS
dn
Discovery / Development / Diagnostics / Delivery

Reprinted with permission
from Drug Discovery News
August 2008
Vol. 4 Issue No. 8

NO GENE LEFT BEHIND

Sophic and Biomax receive funds to complete NCI's Cancer Gene Index Project

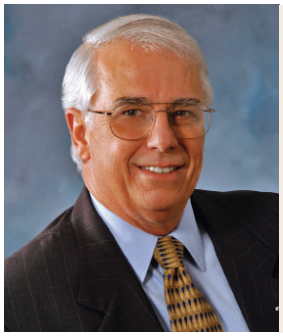
BY AMY SWINDERMAN

EAST FALMOUTH, Mass.—After five years of text mining and manual annotation, Sophic Systems Alliance Inc. has launched the final phase that will complete the National Cancer Institute's (NCI) Cancer Gene Index Project, a highly curated, standardized and computable cancer knowledge base.

Sophic, which teamed up with Biomax Informatics AG in Munich, Germany, to work on the pilot phase of the project in 2004, has received $1.3 million from the NCI to complete the project by May 2009. The NCI has funded the project in whole, or in part, with federal funds. When complete, the cancer community will have access to 6,610 cancer-related genes found in Medline abstracts, with manually annotated gene-disease and gene-compound relationships.

"This knowledge base is the foundation for reliable, accurate cancer research for all types of cancer diseases, clinical trials, biomarkers and much more," says Sophic CEO Pat Blake, project manager for the index. "The NCI has been great to work with, and this has been a real team effort to complete the collaboration. Sophic and Biomax appreciate the opportunity to provide the cancer community with this asset."

The NCI tapped East Falmouth, Mass.-based Sophic to mine 8.8 million Medline abstracts to identify suspect cancer genes, manually verify true cancer genes and manually annotate role codes and evidence codes for 1,000 cancer genes selected by NCI. In order to manage the volume of material to review and organize the complexity of the information for the scientists, Sophic and Biomax CEO Dr. Klaus Heumann developed a "factory assembly line" methodology that allowed the automated text mining results to be fed to the scientific team, which curated and annotated the information in an efficient, quality-controlled workflow process.



"The feedback we have been getting is that this type of high-quality curation is going to become a very strong foundation for all steps of every part of cancer research."

—Pat Blake, Sophic CEO

NCI's Thesaurus was leveraged by Biomax's BioLT Linguistic Tool to create extended dictionaries used to mine the literature. Scientist curators have manually verified all true cancer genes and have used controlled vocabularies to annotate and assign role codes and evidence codes to each gene.

One thousand of the true cancer genes selected by NCI were manually annotated and delivered to NCI in October 2004. From 2005 to 2007, three additional phases of the project were performed, each building on results and lessons learned in previous phases. To date, the scientists have annotated 4,658 cancer genes, which are already publicly available.

"The hallmark of our instructions from the NCI since the beginning has been, 'not to miss anything,'" Blake says. "The feedback we have been getting is that this type of high-quality curation is going to become a very strong foundation for all steps of every part of cancer research."

An index this robust is so valuable, Sophic is in discussions with major hospitals to develop similar libraries for complex diseases, Blake says.

"If you can do something this meaningful for something as complex and terrible as cancer, why not extend that capability to thrombosis or other complex diseases related to cancer?" he says. "This process is going to have the most value for diseases that have huge complexities from a cellular and molecular point of view. Our strategy is to work with doctors and researchers to identify where the tools and annotation factory assembly line methodology developed by Klaus and his team can most effectively be used for research associated with complex disease."