Sophic, Biomax Hone Text-Mining, Curation Tools for Final Phase of Cancer Gene Index

July 7, 2008
By Vivien Marx

Sophic Systems Alliance and Biomax Informatics are sharpening a set of literature-mining and curation tools that they have developed over the last few years as they prepare to complete the Cancer Gene Index, a comprehensive database of associations between genes, diseases, and drug compounds derived from Medline abstracts.

The National Cancer Institute recently awarded the companies a $1.3 million grant to complete the index over the next 12 months using a combination of manual and computational annotation. The project, which kicked off in 2004, has so far curated 4,658 cancer genes with a goal of 6,610 genes by the end of the year.

Approximately 15 people at Sophic and Biomax will use Biomax's BioLT statistical co-occurrence mapping software and BioXM knowledge-management platform for the project. BioLT can be used on any text, but for this project it has been pointed at Medline abstracts, explained Biomax CEO Klaus Heumann. "We sharpened the tool for the requirements of the project," he said.

In addition, the two companies have worked with their collaborators at NCI to develop a workflow that combines computational literature mining with manual curation in a process that Pat Blake, CEO of Sophic Systems Alliance, described as "the factory process of annotation."

The Cancer Gene Index is "a massive project, and you have to have workflow, tools, and discipline around how this is done in order to produce accurate results," Blake said.

Juli Klemm, associate director of Integrative Cancer Research Products and Programs at NCI's Center for Bioinformatics and manager of the Cancer Gene Index project, agreed that the combination of automated text-mining and manual curation is key to the effort.

"That's what makes this such a valuable data set — the combination of both an automated output … but then each of those associations has been manually checked by a scientist," Klemm told *BioInform*.

An Integration Tool

The Cancer Gene Index "provides a structured view of disease/gene and drug/gene associations derived from the biomedical literature that is available to scientists who wish to search for networks of relationships and candidate genes," George Komatsoulis, deputy director of the NCI Center for Biomedical Informatics and Information Technology, told *BioInform* in an e-mail.

"We view the Cancer Gene Index as an integration tool; helping to bind together information related to clinical conditions with the world of molecular biology," he said.

So far, the NCI researchers and their collaborators have completed a pilot phase and three project phases for the project. In the pilot study, which was launched in 2004 and completed in March 2005, the goal was to determine whether it was possible to extract cancer genes and anti-cancer compounds in a reasonable amount of time and annotate them properly, explained Klemm.

That project, which focused on 1,000 genes, was also a rigorous evaluation of the quality of work the companies could deliver, Blake said.

After the pilot, the NCI-Sophic-Biomax team first prioritized the gene list by

"We view the Cancer Gene Index as an integration tool; helping to bind together information related to clinical conditions with the world of molecular biology."

manually annotating those gene-cancer terms and gene-compound associations that had "moderate" sentence counts of around 50 sentences in the scientific literature. The goal of that list was to get "true but less well known information in the literature available in a computable format to help support discovery research," Klemm said. That information "hasn't been cited so many times that it is well known by every cancer researcher."

The focus may have led to some misunderstandings about the dataset. "Sometimes people go in to explore the data set and they type in a commonly known cancer gene as a way of testing it out, but they won't find it because of the order in which we have done the index," Klemm said.

In the next and final phase of the project, however, the focus will be on "high sentence count" genes — those with more than 1,000 references in the scientific literature.

When it is completed, the Cancer Gene Index will be a complete set of annotated cancer genes, both well-known genes and the genes with less known association with cancer, Klemm said.

Hands On, Computer On

The majority of the work over the next twelve months will involve manual annotation of the remaining 1,952 genes in the master list of 6,610 cancer genes, Klemm said.

Sophic and Biomax will use linguistics analysis tools "to look for co-occurences [in sentences] of genes with cancer terms," Klemm said.

"They successfully performed a pilot and then had subsequent option years that were successfully completed and through that they were able to put in place a process and continue to improve that process so that they are quite efficient now at this curation effort," she said.

The companies will deliver data to the NCI every few months for evaluation. It will then be made available to the cancer research community via the Cancer Gene Index.

"We update the dictionaries constantly," said Sophic's Blake. New abstracts, of which there are approximately 7,000 per day, are mined for new terms and concepts and used to refresh the system.

"The automated process pulls out potential genes, which filters things down a lot," Klemm said, noting that the software has so far plowed through 16 million journal article abstracts. "The machine doesn't get it all right, but it is a very good filter," she said.

Although other companies offer co-occurrence mapping tools, applying a generic linguistics tool to a complex area like cancer is "a steep and difficult process," said Biomax's Heumann. "We have focused on cancer and to that end I am not sure there is anyone out there with a generic linguistics co-occurrence tool that could compete in terms of pure accuracy and efficiency with what we are doing."

Heumann explained that there is an "art" to using software to increase the efficiency of manual curation. "Clearly manual labor is rate-limiting. The more effectively you can prepare that, while having the required recall so you don't miss anything and so you don't convolute the results with non-relevant information for precision's sake, those are the two objectives that you cannot master solely by automatic processes."

Once the automated text mining is complete, humans step in. "If they see at least one true, by their reading, sentence that shows an association between a cancer gene and a cancer term, they say, 'That is one we should go in and manually annotate,'" Klemm said.

When the automated process identifies an association that the curators do not deem to be true, that information is counted as a false positive and not annotated, she said, noting that the co-occurrence of two terms in a sentence does not necessarily describe a scientific relationship between the terms.

Assigning the Codes

As the scientific curators read the sentences in the abstracts, they manually assign role codes and evidence codes to the terms. Role codes describe the scientific and semantic association of a given gene with a cancer type or drug term, while evidence codes outline the experimental results that support the annotation.

“An example for a role code could be, ‘chemical or drug affects gene product,’” Klemm said. “An evidence code has items such as, ‘author’s statement,’ or, ‘inferred from computational analysis, or ‘inferred from experiment’ [indicating] how you know that information.”

The evidence code also makes the supporting information “computable,” by mapping it to defined vocabularies, Heumann said.

Specifically, the project draws on the NCI Thesaurus, a controlled vocabulary that assigns a “concept code” to a given term that acts as “a stabilizing feature” for terms that can have many synonyms, Klemm said

“When Sophic and Biomax do their automated mining, they can use all [possible] synonyms when they search the literature, but those all map back to the relevant concept code,” Klemm said.

In addition to the BioLT literature-mining tool, the project partners will use Biomax’s BioXM platform for integrating and managing the highly curated content, making it the “backbone” of the project, Blake said.

Heumann added that BioXM will allow researchers to take a lateral view, connecting genes to disease and genes to compounds and also to a whole range of other information such as pathways, and gene expression data. “You can now connect these things, ideally through BioXM, and navigate across silos,” he said.

Sophic and Biomax have a longstanding relationship. Sophic, a US government contractor and the lead integrator for this project, was originally founded to partner with Biomax Informatics and distribute its products in the US. The project team is thinking about ways to make transparent how the information in the Cancer Gene Index was generated. One way, said Klemm, is a paper in a peer-reviewed journal.

“Having a peer-reviewed article with very detailed descriptions of analyses and statistics on the data will be an important resource to the community to go along with the data set,” she said.