EXECUTIVE SUMMARY
Pharmaceutical research is focused on understanding the role our immune systems play in the prevention, progression, and potential treatments for cancer. The complexity of the interactions and processes of the immune system make understanding the underlying biology difficult and result in a very diverse collection of studies in the literature. To help researchers address these challenges, Elsevier has developed novel text-mining tools to find and aggregate data on cellular processes into a new data source for use with Pathway Studio.
Introduction

Following on the successful treatment of melanoma with novel immunotherapies, the search for oncological therapeutics has lately become increasingly focused on immune cells. As the research trend started heating up, an Elsevier team saw an opportunity to contribute a resource to the field.

“It’s important to know what happens in the tumor microenvironment,” says Dr. Maria Shkrob, Project Manager at Elsevier. “There is an ongoing battle: immune cells try to stop tumors from spreading, but malignant tumors develop the means to locally suppress the immune system.”
The tumor microenvironment consists of tumor cells, cells surrounding a tumor, cells inside of a tumor, and various molecules that are secreted by or exposed outside of those cells. Interactions among these players are important for disease progression and therefore of interest to cancer researchers. “We know lots of bits and pieces of information about those interactions, but the problem is that they are scattered across hundreds of thousands of publications,” Shkrob says. “Humans can’t read that much to put everything together, but machines can.”

Dr. Shkrob’s group works on automatic extraction of information from biomedical publications to populate Elsevier’s Pathway Studio knowledgebase, which now contains over 5 million connections among proteins, chemicals, and diseases, derived from literature. Last year her team undertook a project to expand the knowledgebase coverage with information about cells. The goal was to enhance support for users focused on immunology—cancer immunology, in particular.

To assemble the most useful collection of data for immunology research, Shkrob says the team first set out to determine what information about cells interests scientific researchers. They discovered common questions such as: Which cells play an active role in disease initiation or progression? Which cells are affected by disease? How do they change their phenotype? Does the number of cells change? What effects do proteins and drugs have on cells? Which proteins are secreted from cells? And which are exposed on their surfaces?

**Figure 1.** Exponential growth of MedLine publications mentioning cancer immunology.
Once the scope was defined, the team needed to tailor Elsevier’s natural language processing engine to capture the facts about cells. The engine splits publications into sentences, recognizes biological concepts within a sentence, and then uses linguistic algorithms to extract the connection between these concepts.

The biggest challenge to the project came from the lack of naming conventions for immune cells in the current literature. Shkrob explains: “Along with more or less typical names, such as ‘regulatory T-cell,’ ‘T regulatory cell’ or ‘Treg,’ immunologists often define a cell type using certain surface proteins, for example ‘CD3+ CD4+ CD25+ T cell,’ or ‘CD25pos CD4pos T-lymphocyte.’ All these terms refer to the same cell type. Imagine what problems this variety creates, when you need to find relevant information.”

Shkrob says the approach Elsevier scientists have used previously to group synonyms, such as “acute myeloblastic leukaemia” and “acute nonlymphocytic leukemia,” was insufficient for deciphering immunology semantics. They had to develop an additional algorithm that would take into account all combinations of cell epitopes.

The improved engine was used to process more than 25 million abstracts and 3.5 million full-text publications. It captured more than 800,000 unique connections, increasing the total number of relationships in the Pathway Studio knowledgebase by more than 15 percent.

**Figure 2.** Pathway Studio helps researchers connect information about cells, proteins, diseases, small molecules, and other observations from the literature.
Dr. Shkrob points to two key ways the biomedical research community will benefit from access to the cell-centered data models her team has built. First, these literature summaries will enable researchers to get answers to key research questions significantly faster. “Even if your question sounds very simple, such as, ‘What cells are involved in my disease of interest?’ it might be quite hard to answer. You might read reviews, but you can’t be sure that the reviewer actually addressed every relevant piece of information from the literature,” Shkrob says. “You can do a search yourself using keywords, but you might find thousands of papers to read, and still be missing some because of the naming inconsistencies.”

Shkrob says that for a more complex question such as, “Can breast cancer cells promote their own growth—secreting factors that affect immune cells in the microenvironment?” a user might employ Pathway Studio to find the proteins secreted from breast carcinoma cells, find which immune cells regulate breast carcinoma progression, and find out how proteins secreted from breast carcinoma cells affect those immune cells, all much more quickly than via a conventional literature search.

Figure 3. Breast carcinoma cells often secrete proteins that can promote or inhibit cancer progression in surrounding cells. The data used to create this representation came from over 200 different publications, enabled by Pathway Studio.
In a second application, she says, Pathway Studio powered by the new knowledgebase can be used to better interpret experimental data such as gene expression or next-generation sequencing studies.

“One of the advantages of automated approaches is their flexibility,” says Shkrob. “It is easier to keep up with the pace of the field and to meet researchers’ needs. We can extend the data model and easily tweak our definitions of cells, and then quickly reprocess the literature to create an updated version of the database.”

As immunology-related research explodes and advances in research methodology enable simultaneous detection of dozens of different proteins on a cell surface, cell naming will undoubtedly get more complicated. Shkrob says the new approach gives Elsevier’s Pathway Studio the ability to deal with the ever-increasing complexity of biological research.
LEARN MORE

To request information or a product demonstration, please visit elsevier.com/pathwaystudio or email us at pathwaystudio@elsevier.com.