Using Linguamatics NLP For Target Prioritization At Pfizer

This case study describes how the Linguamatics NLP platform has been used to capture valuable information from life science literature, saving time and increasing productivity

Challenge

The Pfizer Research Technology Center (RTC) in Cambridge, MA, houses cutting-edge research groups focused on RNAi therapeutics, Systems Biology and Biological Profiling, Regenerative Medicine and Computational Sciences. Within the RNAi therapeutics area, the challenge facing scientists was to survey the literature for studies reporting gene or protein behaviour consistent with RNAi as a therapeutic approach. Specifically researchers needed to extract knowledge from text demonstrating that reduced activity or abundance of a target ameliorated disease phenotypes in vivo or in vitro. As RNAi is a relatively new experimental approach, no databases or public repositories contained the necessary information, thereby necessitating a thorough literature review. Furthermore, the information needed to be gathered in a matter of weeks, ruling out a manual literature review as a viable approach.

Solution

Teams at Pfizer are using Linguamatics NLP to provide an agile text mining platform for querying large external document collections, extracting relevant facts, relationships and numerical information. Semantic search capabilities were enhanced by plugging in domain-specific ontologies and thesauri to enable queries to automatically find relevant synonyms or search for entire classes of items. Results describing the relationships between targets and disease were grouped by target and presented as structured tables that included the particular match found for a concept or its preferred term, to enable easy integration with structured knowledge sources.

Querying targeted publicly available document sources such as MEDLINE® and other literature databases to categorize candidate targets. Researchers defined several categories of queries using keywords and linguistic expressions, exploiting the Linguamatics platform's powerful NLP querying capabilities, including parts of speech, proximity controls and mapping to preferred terms, to control the precision of results. Queries that used multiple linguistic patterns to retrieve related information were combined into multi-queries for efficient compilation of results.

> *"Extracting information about proteins that share a specific set of molecular behaviors and how they relate to disease is impossible to accomplish in a short time frame with typical search engines such as PubMed or Google. [Linguamatics] enabled us to rapidly build custom vocabularies and apply them to generate high-quality results that were easily integrated with other data sources, to present scientists with a single comprehensive overview of the best targets for a therapeutic approach."*
>
> *— Phoebe Roberts, Senior Principal Scientist, Pfizer*

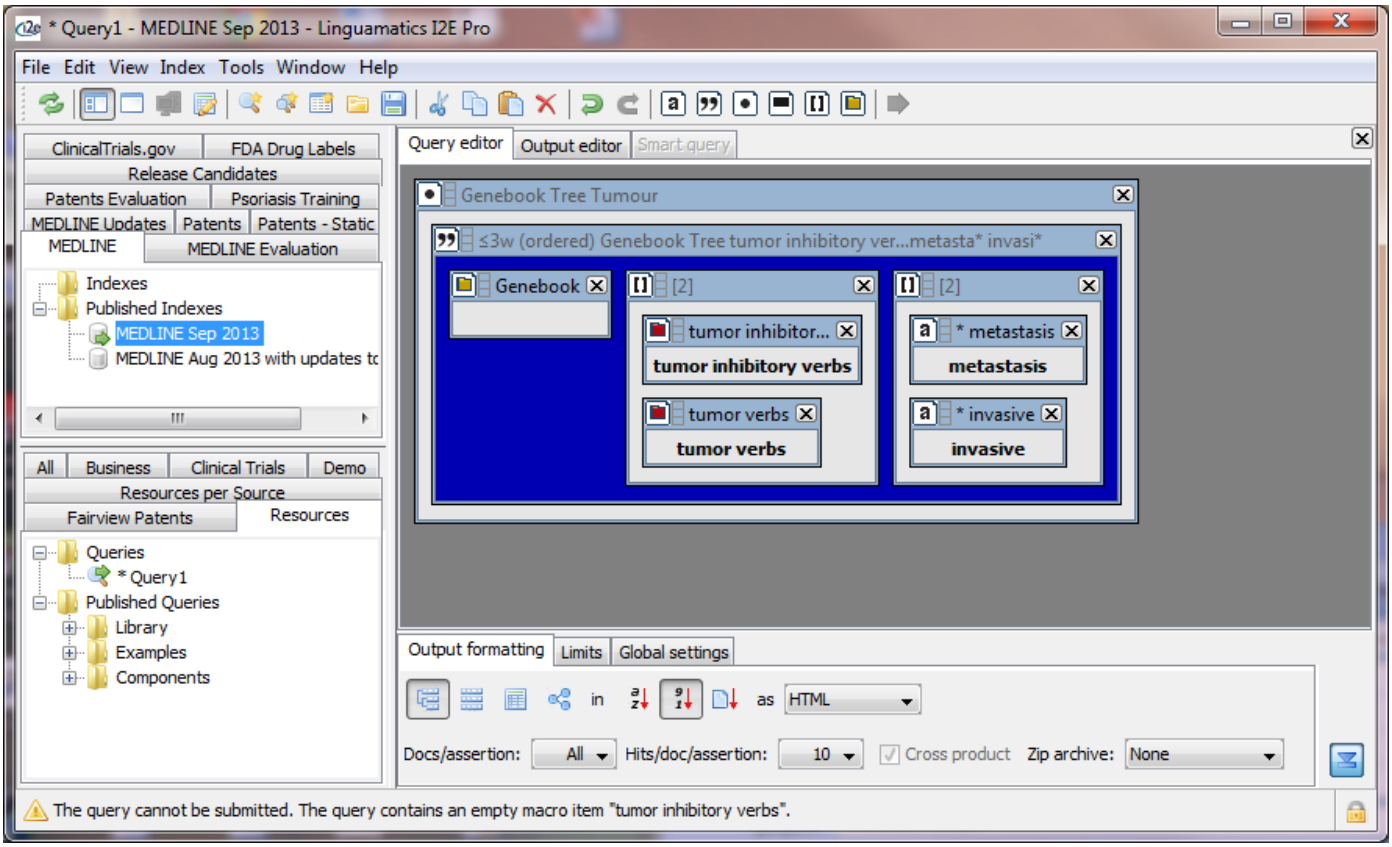
The NLP platform was also used to discover synonyms to build a vocabulary containing terms that were representative of usage in the documents of interest. It was known that RNAi technology using siRNA is capable of sequence-specific gene silencing, and that a particular RNAi therapeutic for inhibiting the VEGF gene was entering Phase III clinical trials, indicating that VEGF was well-studied and could be used as a positive control for retrieving relevant literature. Reports of VEGF knockdown by RNAi were used to build the vocabulary that was then plugged into the platform. Ontologies were also used to cluster results based on stimulatory and inhibitory relationships between candidate targets and disease.

Structured results were presented as tables to allow easy curation by Pfizer scientists. The curated tables were combined with output from other external and internal target knowledge databases to present a comprehensive picture of each target, including whether it had been targeted by other therapeutic modalities, whether in vivo data were available, or if the target had been associated with human disease. The results mined from text, combined with data from at least ten other sources, allowed scientists to filter results by various criteria to come up with a short list of targets. Links to original sources generated by the platform allowed scientists to pursue tantalizing leads.

Examples of specific types of query include:

- **Promoting queries** that suggest a protein might be amenable to RNAi, for example: “Finally, KLF6 silencing caused p53 upregulation and inhibited Bcl-xL expression, to induce cell death by apoptosis”; “Blockade of MIF bioactivity successfully inhibited tumor cell growth in vivo and in vitro.”
- **Demoting queries** where experimental results suggest that the RNAi approach would be detrimental, causing effects such as loss of heterozygosity or prevention of apoptosis: “Deletion of NEMO/IKK(gamma) in liver parenchymal cells causes steatohepatitis and hepatocellular carcinoma”; “Loss of heterozygosity of M6P/IGF2R gene is an early event in the development of prostate cancer.”
- **Informing queries** where query results add information about the role of the target in cancer and other indications.

Figure 1 Shows A Query For Retrieving Sentences About Genes Or Proteins Implicated In Cancer



Summary

The flexible text mining approach described here shows how researchers can address questions in a highly specific domain by identifying and implementing relevant terms. Querying is user-controlled within a user-defined context.

The speed of search provides rapid access to new insights much faster than by traditional methods.

Powerful underlying linguistics enable queries to be defined to identify relevant relationships that answer specific questions, rather than just finding documents.

In addition, searches are reproducible. Queries can be saved, re-run on alternative data sources, and easily modified to compare alternative search strategies or to apply best practice to related problems.

CONTACT US

+44 (0)1223 651 910 (U.K.) | +1 617 674 3256 (U.S.)

nlp@iqvia.com

linguamatics.com

©2020. All rights reserved. IQVIA® is a registered trademark of IQVIA Inc. in the United States, the European Union, and various other countries. 08.2020.RWS