



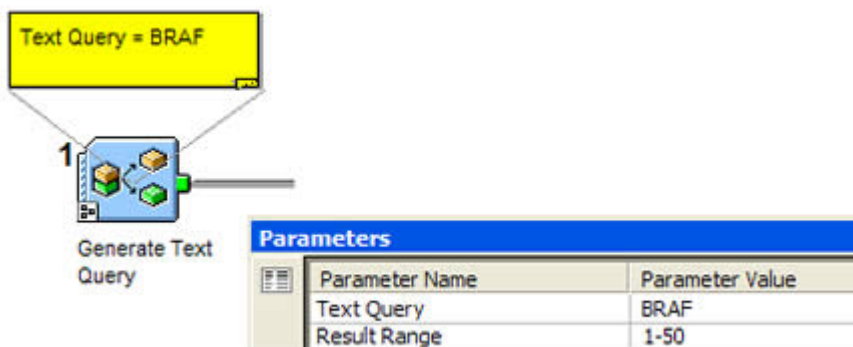
Information Extraction from Text Documents using Pipeline Pilot

With so much text-based information currently available, and more content becoming available all the time, it is ever more vital to have effective, efficient means to extract information from relevant documents. It is simply not possible for you to read all the documents. On the other hand, to completely automate the processing of documents can be dangerous, since the machine process can never match the human ability for language comprehension and reasoning. The solution is an approach that utilizes the respective strengths of human and machine processing such that the machine component takes care of the repetitive, automatable tasks so that the key documents can be presented to the user in a form that facilitates review and decision-making.

In the following use-case, we are interested in determining the relationship of a gene (BRAF) with various forms of cancer. We don't want to have to perform multiple searches for BRAF and each form of cancer. Instead we want an efficient, repeatable approach that takes a gene query and relates the retrieved documents with cancer terms.

Create the Text Query

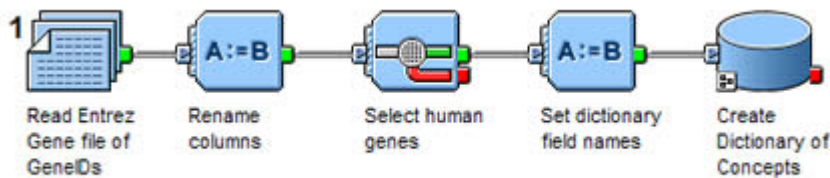
The first step is to create a text query. For a single query, start with the Generate Text Query component and set the query to the gene "BRAF" and ask for the first 50 hits:



Expand the Text Query with Synonyms

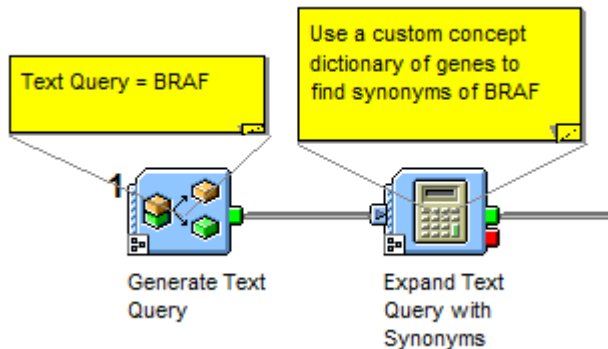
Genes, as with many other concepts that appear in text, have a number of synonyms. To ensure that you perform a comprehensive search, it is important to expand your query to include any synonymous terms. PubMed does this automatically, but if you will be searching other data sources, or to ensure the most comprehensive searching, you can expand your query using a dictionary of terms. To do this you can use any concept dictionaries that come prepackaged with Pipeline Pilot (e.g., MeSH, the MeSH dictionary from the National Library of Medicine) or you can create your own custom concept dictionaries.

The following protocol shows how a copy of the Entrez Gene file containing gene IDs and synonyms, downloaded from NCBI, can be used to create a custom dictionary of gene synonyms:



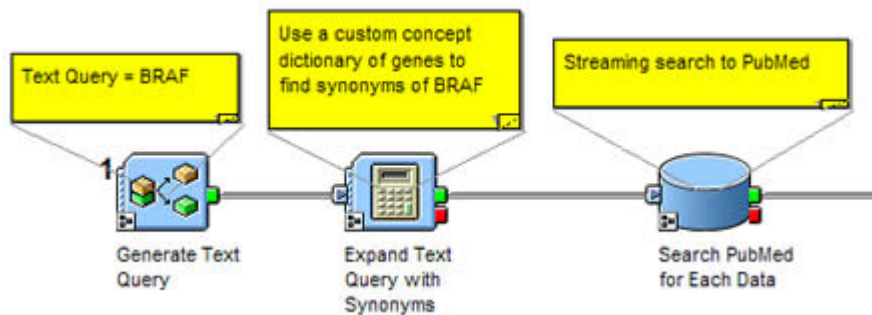
Note that to keep this up to date you could include the download step as part of the protocol, and set the job to run every night.

Once the dictionary is created, use the "Expand Text Query with Synonyms" component to expand the BRAF query:



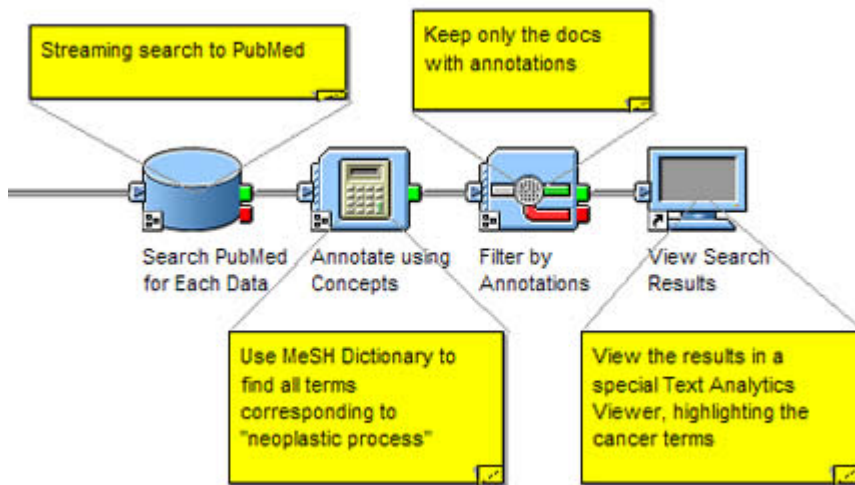
Performing the Search

The next step is to perform the search - in this example we will search PubMed. Searching can occur in one of two ways, either by starting with a search component (see the Pub Med Trend Analysis case study), or by using streaming data to search with. In this example, we use the latter approach, which would allow us to not just search a single gene (BRAF) but we could instead do an individual search for a list of genes of interest.



Extracting Information from the Documents

Once the documents have been retrieved the next step is to find the relationships with types of cancer. To do this we Annotate using Concepts, to find all occurrences of "cancer" terms from the MeSH dictionary in the retrieved documents. We then keep only those documents that contain such terms. In this simple example, we will just display the results but in more advanced use cases we could do further processing to count the number of documents by types of cancer, filter for specific forms of cancer, calculate trends and correlations between BRAF and various forms of cancer, or any number of other types of analysis, according to our research interest.



Viewing the Output

The following graphic shows the output of this protocol. In the header of the webpage output, the online source (PubMed), the result range (1 - 50), the total number of matching documents (734) and the expanded text query are shown. The first matching document is then shown, with the query term highlighted in yellow and the matching MeSH cancer terms (e.g., carcinoma) highlighted in blue.

PubMed: Results 1 to 50 of 734 for B-raf 1 OR BRAF1 OR RAFB1 OR BRAF

- BRAF** Is a Therapeutic Target in Aggressive Thyroid Carcinoma.

Clin Cancer Res 2006; Salvatore, Giuliana; De Falco, Valentina; Salerno, Paolo; Nappi, Tito Claudio; Pepe, Stefano; Troncone, Giancarlo; Carlomagno, Francesca; Mellillo, Rosa Marina; Wilhelm, Scott M; Santoro, Massimo

Authors' Affiliations: Istituto di Endocrinologia ed Oncologia Sperimentale del Consiglio Nazionale delle Ricerche, c/o Dipartimento di Biologia e Patologia Cellulare e Molecolare.

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=retrieve&db=pubmed&doct=Citation&list_uids=16533790

... levels. CONCLUSIONS: **BRAF** provides signals crucial for proliferation of thyroid carcinoma cells spontaneously harboring the (V600E)**BRAF** mutation and, therefore, **BRAF** suppression might have therapeutic potential in (V600E)**BRAF**-positive thyroid cancer. ... PURPOSE: Oncogenic conversion of **BRAF** occurs in approximately 44% of papillary thyroid carcinomas and 24% of anaplastic thyroid carcinomas. In papillary thyroid carcinomas, this mutation is associated with an unfavorable clinicopathologic outcome. Our aim was to exploit **BRAF** as a potential therapeutic target for thyroid carcinoma. EXPERIMENTAL DESIGN: We used RNA interference to evaluate the effect of **BRAF** knockdown in the human anaplastic thyroid carcinoma cell lines FRO and ARO carrying the **BRAF** V600E ((V600E)**BRAF**) mutation. We also exploited the effect of BAY 43-9006 [N-(3-trifluoromethyl-4-chlorophenyl)-N'-(4-(2-methylcarbamoyl)...

A second result shows that BRAF is also associated with melanoma:

- Regulation of iNOS by the p44/42 mitogen-activated protein kinase pathway in human melanoma.

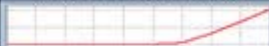
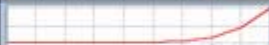
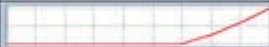
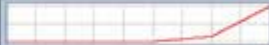
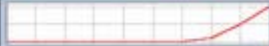
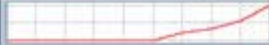
Oncogene 2006; Ellenhorst, J A; Ekmekecioglu, S; Johnson, M K; Cooke, C P; Johnson, M M; Grimm, E A

1The Department of Experimental Therapeutics, The University of Texas, MD Anderson Cancer Center, Houston, TX, USA.

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=retrieve&db=pubmed&doct=Citation&list_uids=16474847

... pathway. On the basis of our data showing that melanoma iNOS expression predicts shortened patient survival, we formulated the hypothesis that activating mutations of NRAS or **BRAF**, which lead to constitutive activation of the p44/42 MAPK pathway, drive iNOS expression in human melanoma. In the present study, we have shown that inhibition of melanoma iNOS activity by S-methylisothiourea leads to decreased cell proliferation, confirming the importance of iNOS activity for melanoma cell growth. Regulation of melanoma iNOS expression by the p44/42 MAPK pathway was demonstrated by inhibition of the pathway by U0126, and by **BRAF** RNA interference. To explore this regulatory pathway in human tissue, 20 melanoma tumors were examined for NRAS and **BRAF** mutations, immunohistochemical evidence of ERK phosphorylation, and iNOS expression. A significant association was found among these three features. We ...

As well as viewing the individual documents, you can collate them into a table showing the count of documents and the recent publication frequency to reveal what forms of cancer are most commonly, or most rarely, associated with BRAF, and when those documents were published:

| Cancer indications (MeSH concept) | No. Articles (% Total) | Cumulative frequency of articles (<=1996 to 2005) |
|-----------------------------------|----------------------------|---|
| Melanoma | 69 (18.3%) |  |
| Carcinoma | 42 (11.1%) |  |
| Neoplasm Metastasis | 20 (5.3%) |  |
| Colorectal Cancer | 14 (3.7%) |  |
| Carcinoma, Papillary | 10 (2.7%) |  |
| Adenoma | 9 (2.4%) |  |

This relatively simple protocol brings together the collective resources of Entrez Gene, PubMed and MeSH, representing thousands of hours of research by hundreds of individual scientists, to retrieve and highlight the most important documents for you to review and analyze. The deeper your research questions go, the more evident the power of pipelining your text analyses becomes

Copyright © 2001-2008 Accelrys Software Inc.

[Careers](#) | [Legal / Terms of Use](#) | [Contact us](#)