Text mining workflow for extraction of paragraphs from full articles describing drug-gene interactions to support Onco KEM software platform for personalized treatments

Fanny Perraudaud, David Morley, Mohammad Afshar, Mariana Guergova-Kuras
Ariana Pharmaceuticals, 28 Rue du Docteur Finlay, 75015 Paris France, +33(0) 1 44 37 17 00, d.morley@arianapharma.com

Introduction

Synthetically presenting key information about drug-gene interactions to support personalized treatment decisions is a challenge and requires databases created by curation of full scientific articles. Databases, such as Comparative Toxicogenomics Database (CTD) or Drug-Gene Interaction (DGI), provide a brief sentence for this purpose, supported by the associated PubMed ID, which in numerous cases do not allow quick evaluation of the relevant content. We describe here a text-mining workflow (Onco KEM® Builder) for automatic screening of full text articles aimed at extracting complete paragraphs of scientific evidence on drug-gene interactions and creating a database to be integrated in a software platform for personalized treatments in oncology.

Pipeline for data processing

1. Corpus of scientific literature (currently papers from PubMed®)
2. Onco KEM® Builder extracts the relevant knowledge: the most appropriate paragraphs explaining the drug-gene interactions
3. The medical oncologists use the Onco KEM® database to support personalized treatment decisions

Query Example

Query: « Ofatumumab binds to the CD20 antigen where upon it induces cell lysis » + PubMedID : 22150234
Extracted paragraph: « Ofatumumab is a human IgG antibody that binds to a unique, more membrane proximal epitope of the CD20 antigen. Pre-clinical studies have shown ofatumumab to have similar antibody-dependent cellular cytotoxicity (ADCC) and improved CMC when compared to rituximab. Ofatumumab induces prolonged B-cell depletion when compared to rituximab and has been shown to slow lymphoma tumour cell growth in xenograft models. Ofatumumab was recently approved by the Food and Drug Administration (FDA) for the treatment of fludarabine and alemtuzumab refractory chronic lymphocytic leukaemia. »

Methods - Workflow steps

Preliminary work

The corpus consisted of 56 PubMed articles downloaded with PaperToolBox® to extract paragraphs related to 115 sentences from CTD describing drug-gene interactions for 34 cancer related drugs. The customized PDFMiner python library® extracted the text from the PDF files of the full articles. The figures, tables and references of the articles were not extracted avoiding false negative paragraphs extraction. The keywords constituting the sentence were enriched with synonyms for both drugs and genes. The synonyms of the drugs were extracted from the CTD database. For the synonyms of the genes, the "gene_info" file downloaded from the FTP of the NCBi was used.

Text Mining Tool

A KNIME workflow based on dictionary tagger and bag-of-words model was used to score each paragraph according to the number of occurrences of the keywords. Custom R script reassembled the paragraphs, if the paragraphs with the highest score were incomplete after converting PDF to text. The assembly of the paragraphs is based on a sentence boundary detection algorithm. If the end (resp. start) of a paragraph was not considered as the end (resp. start) of a sentence, the algorithm looked for the next paragraph not starting (resp. ending) with a complex regular expression representing the start (resp. end) of a sentence.

Database building

Database was completed with the paragraphs selected as the most representative by the curator. If no paragraph was considered as relevant among the ranked ones, the abstract of the article is chosen.

Result Table

<table>
<thead>
<tr>
<th>Steps</th>
<th>Most appropriate paragraph ranked n°1</th>
<th>Most appropriate paragraph ranked n°2,n°3,n°4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before paragraphs assembly</td>
<td>74</td>
<td>41</td>
</tr>
<tr>
<td>A single paragraph ranked as highest</td>
<td>93 (accuracy = 80.9%)</td>
<td>22</td>
</tr>
<tr>
<td>Up to three paragraphs ranked highest</td>
<td>53</td>
<td>40</td>
</tr>
<tr>
<td>Most appropriate paragraph ranked n°2</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>Most appropriate paragraph ranked n°3,n°4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results

In 93 cases out of 115 sentences, the most appropriate paragraph in the article to describe the drug-gene interaction was ranked number one by the workflow. For 53 sentences a single paragraph was ranked as highest; and in the remaining 40 cases up to three paragraphs were equal in ranking. For the 22 other sentences, the most appropriate paragraphs were ranked no lower than fourth.

In summary, our text mining workflow reduces knowledge base curation time by selecting the most appropriate paragraph with an accuracy of 80.9%. On average, articles can be curated five times faster using the workflow compared to manual curation of full articles.

The need to use full article

For 56 sentences, the extracted paragraphs were not from the abstracts of the articles but from the core text, showing that databases relying only on article abstracts may not provide sufficient information for clinical decision making.

Conclusion

Our Onco KEM® Builder text-mining workflow can generate a knowledge base of complete paragraphs describing drug-gene interactions to support clinicians in personalized treatment decisions for patients.

References