Ongenaert and Dehaspe *BMC Bioinformatics* 2010, **11**(Suppl 5):O5
http://www.biomedcentral.com/1471-2105/11/S5/O5

BMC Bioinformatics

ORAL PRESENTATION Open Access

Integrating automated literature searches and text mining in biomarker discovery

Maté Ongenaert*, Luc Dehaspe

From Workshop on Advances in Bio Text Mining
Ghent, Belgium. 10-11 May 2010

Background

Epigenetics, and more specifically DNA methylation is a fast evolving research area. In almost every cancer type, each month new publications confirm the differentiated regulation of specific genes due to methylation and mention the discovery of novel methylation markers. The last decade, high-throughput methodologies are frequently used in the discovery of such methylation biomarkers. Examples of such analyses are re-expression experiments (using the demethylating agent 5-Aza-2′-Deoxycytidine, followed by expression micro-array analysis); CpG microarrays such as the Illumina HumanMethylation27 BeadChip and large scale bisulfite sequencing.

In order to evaluate and to prioritize possible methylation biomarkers, a literature search is a good starting point. However, manual searches are time-consuming (as hundreds of genes are to be searched, taking all their aliases into account) and the summarization of the found references is a real challenge. Therefore, it would be extremely useful to have an annotated, reviewed, sorted and summarized overview of all available data, published in methylation research in cancer.

Results

In a first stage, an automated literature retrieval and annotation tool was created, code-named GoldMine. This web-based application allows entering a list of genes, keywords and highlighting terms. Of the genes, all aliases are used to search PubMed abstracts, in combination with the keywords. The gene aliases, the keywords and the highlighting terms are highlighted in different colors as well as sentences with both a gene alias and a keyword. Abstracts are presented with decreasing scores that are assigned.

Based on this framework, a cancer methylation database is created: PubMeth (as shown in Figure 1). PubMeth [1] is a cancer methylation database that contains genes that are reported to be methylated in various cancer types. A query can be based either on genes (to check in which cancer types the genes are reported as being methylated) or on cancer types (which genes are reported to be methylated in the cancer (sub) types of interest).

More recently, in the context of the SBO project on Functional Peptidomics, the MouseMining tool was developed to further exploit PubMeth results and comparable literature summary data by combining them with experimental data. In a prototypical application, MouseMining was used to correlate statistics on the co-occurrence of anatomic categories and disease names to the expression profile of candidate biomarkers.

Conclusions

The generated methylation database in cancer is freely accessible at http://www.pubmeth.org. PubMeth is based on text mining of Medline/PubMed abstracts, combined with manual reading and annotation of preselected abstracts. The text mining approach results in increased speed and selectivity (as for instance many different aliases of a gene are searched at once), while the manual screening significantly raises the specificity and quality of the database. The summarized overview of the results is very useful in case more genes or cancer types are searched at the same time.

Published: 6 October 2010

* Correspondence: Mate.Ongenaert@OncoMethylome.com
OncoMethylome Sciences, 4000 Liege, Belgium

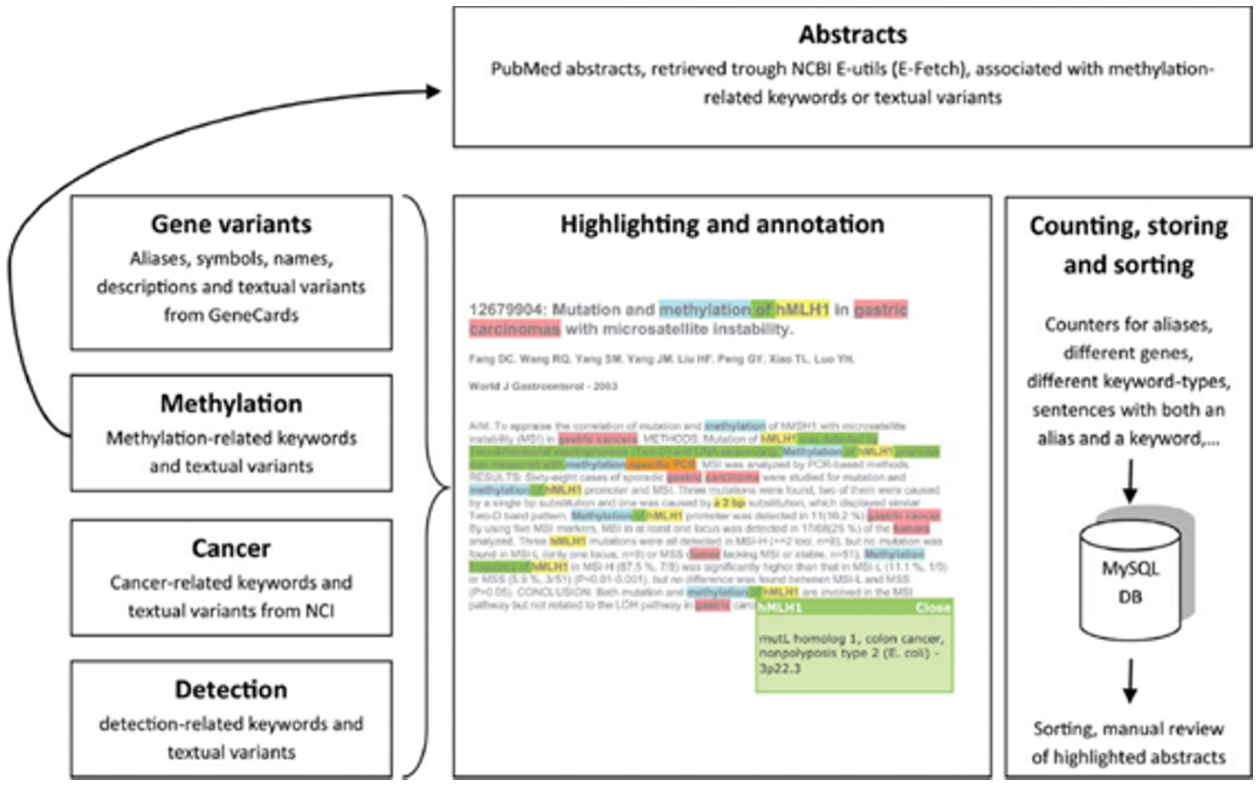

Figure 1 Scheme that illustrates the initial filling up of database using text mining. Aliases of genes and different keyword lists (methylation, cancer and detection-related) are highlighted in the abstract. At the same time, different parameters are counted and stored in a MySQL relational database. Afterwards, the data is ranked and manually reviewed.

Reference

1. Ongenaert M, Van Neste L, De Meyer T, Menschaert G, Bekaert S, Van Criekinge W: **PubMeth: a cancer methylation database combining text mining and expert annotation.** *Nucleic Acids Res* 2008, **36**:D842-D846.

doi:10.1186/1471-2105-11-S5-O5
Cite this article as: Ongenaert and Dehaspe: **Integrating automated literature searches and text mining in biomarker discovery.** *BMC Bioinformatics* 2010 **11**(Suppl 5):O5.