SciBite
an ELSEVIER company

Use case

Biomarker discovery in literature

SciBite
an ELSEVIER company

Biomarker discovery in literature

The identification and application of biomarkers in basic and clinical research is almost a mandatory process in any productive pipeline of a biopharmaceutical organisation. Validated biomarkers play a crucial role in the prediction of clinical outcome, support the translation from candidate discovery to successful clinical treatment. The process to discover and validate new biomarkers depends on effective methodologies often calling on text mining approaches to extract insight from biomedical literature. The following white paper evaluates SciBite's capabilities in identifying new gene biomarkers in Breast Cancer against a published methodology.

Given the wealth of information available in biomedical literature, an important thing is to be knowledgeable of all the existing biomarkers and also other biomolecules that may be suitable as new biomarkers. SciBite's TERMite and TExpress products provide a powerful and effective text mining solution that can identify and extract new potential biomarker leads from scientific text.

To demonstrate the effectiveness of the SciBite biomarker discovery methodologies, we have undertaken a comparison with the published research of Jordan et al 2014 (Journal of Clinical Bioinformatics 2014, 4:13)[1] entitled "Semi-automated literature mining to identify putative biomarkers of disease from multiple biofluids".

Methodology

The Jordan paper outlines a comprehensive strategy for searching the Medline database for biomarkers associated with a given disease (breast cancer) and presence in different bodily fluids such as blood, sweat etc. While a biomarker can be any measurable biomolecule, we will focus on genes as biomarkers to be comparable with the results of Jordan et al.

The authors used the open source "Abner" software for gene name recognition. To demonstrate the greater performance of TERMite, we repeated and then enhanced the experiments shown in the paper.

A corpus of abstracts related to breast cancer was generated from a simple PubMed query using the remote query service of TERMite.

'breast cancer AND (biofluid)',
(e.g. breast cancer AND blood)

This xml was then used for mining. We applied a three-part approach to search for breast cancer biomarkers as detailed below.

1a Known Biomarker Identification. An important methodology is to identify those biomolecules already established as breast cancer biomarkers. SciBite's TExpress is specifically designed to recognise linguistic patterns within text. The pattern to identify a biomarker would look like:

```
:(GENE|BIOCHEM|PROTYP):{0,3}:MARKER:{0,3}:(INDICATION)
```

This translates to "find any gene, biochemical or protein class within 3 words of a marker term within 3 words of an indication". Such a pattern is then applied to a large corpus of text and any text fragments matching this pattern are returned. Thus, the phrase "BRCA1 expression is therefore a good biomarker for breast cancer" would be identified. As part of this work we developed an extensive set of biomarker-related terms and phrases such as 'prognostic', 'marker of', 'useful indicator', 'predicts' etc. that cover many different ways of conveying such facts. This was used to create our MARKER VOCab used in this methodology. SciBite has a collection of over 50 highly enriched vocabularies (VOCabs) containing over 20 million scientific terms and their related synonyms. Unlike most publically available ontologies, these are specifically designed for text mining

[1] http://jclinbioinformatics.biomedcentral.com/articles/10.1186/2043-9113-4-13

1b Potential Biomarker Identification. In addition to identifying known biomarkers, scientists will also be interested in biomolecules who's properties seem to correlate in some way with a disease state, though are yet to be fully referred to as a biomarker. Such an example may be the increased expression of a gene as a particular disease progresses. In order to find such mentions we need to look for more loosely associated correlations to look for phrases hinting at some connection between a biomolecule and a disease. For this we developed the MARKERPLUS VOCab that contains phrases such as "correlates with", "detects", "altered expression", "associated with" and many more.

2 Gene Mentions. Finally, it may be the case that biomolecules are mentioned in relation to breast cancer, but not in context that can be identified by a simple pattern. For example, a document may discuss the relationship between a molecule and disease but never provide a summary statement as to the utility of the molecule as an indicator. Thus, we can perform a simple search using TERMite simply to list out these biomolecules. While this will give the maximum recall, this will also likely identify molecules that are not candidates as biomarkers. As shown below, the different query strategies offer a different balance of precision and recall.

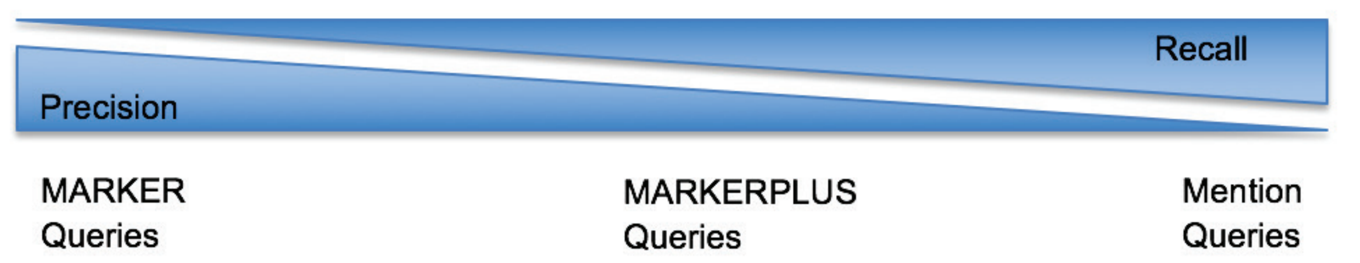


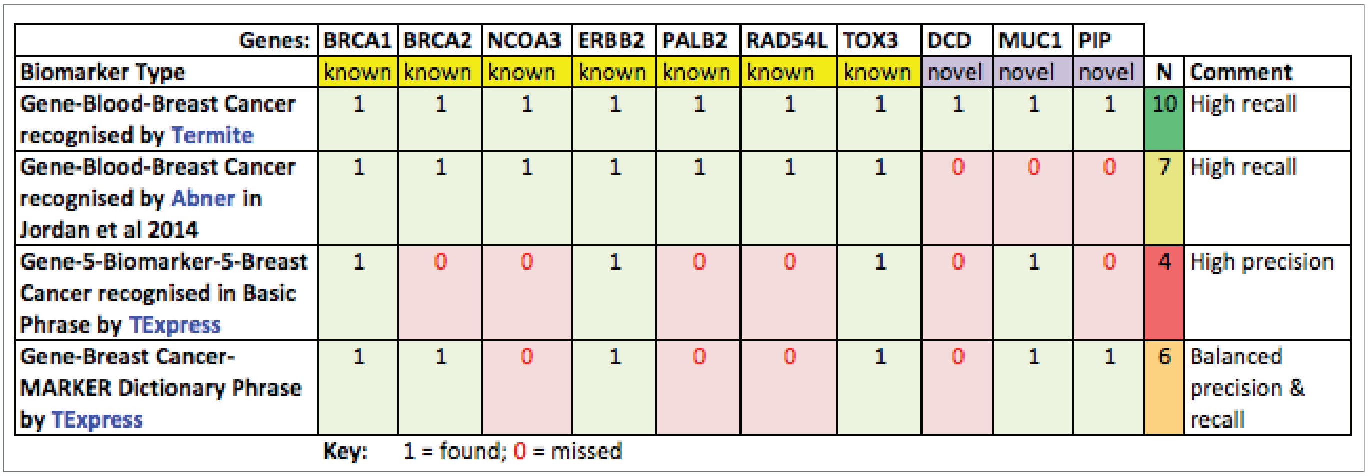
Figure 1: Different query strategies offer a different balance of precision and recall

With this three-part approach, users can be assured they are getting all possible data and are then able to filter accordingly.

At this point, it should be stated that Jordan et al only used the Gene Mentions approach and not either of the linguistic patterns of 1a or 1b discussed above. In total we analysed 20,372 Medline articles with a date cut-off of 29 April 2013 to retain compatibility with Jordan et al.

Results

As stated above, Jordan et al only performed co-occurrence queries (2. Gene Mentions), the Abner software does not have the capability to search for the biomarker linguistic patterns. Thus, the most accurate comparison is on the gene mention data alone.



Genes:	BRCA1	BRCA2	NCOA3	ERBB2	PALB2	RAD54L	TOX3	DCD	MUC1	PIP		
Biomarker Type	known	known	known	known	known	known	known	novel	novel	novel	N	Comment
Gene-Blood-Breast Cancer recognised by Termite	1	1	1	1	1	1	1	1	1	1	10	High recall
Gene-Blood-Breast Cancer recognised by Abner in Jordan et al 2014	1	1	1	1	1	1	1	0	0	0	7	High recall
Gene-5-Biomarker-5-Breast Cancer recognised in Basic Phrase by TExpress	1	0	0	1	0	0	1	0	1	0	4	High precision
Gene-Breast Cancer-MARKER Dictionary Phrase by TExpress	1	1	0	1	0	0	1	0	1	1	6	Balanced precision & recall

Key: 1 = found; 0 = missed

Table 1: Comparing the results from SciBite against Jordan et al.

As shown in **ROW 1** of Table 1, all known biomarker genes mentioned in blood for breast cancer as reported by Jordan et al were identified by TERMite but it also finds **three additional biomarker genes** (DCD, MUC1 and PIP) not reported by Jordan et al (Jordan results are shown in **ROW 2**). This is a clear demonstration of the coverage of SciBite's highly curated VOCabs over those in the public domain and relative ease in repeating the experiment.

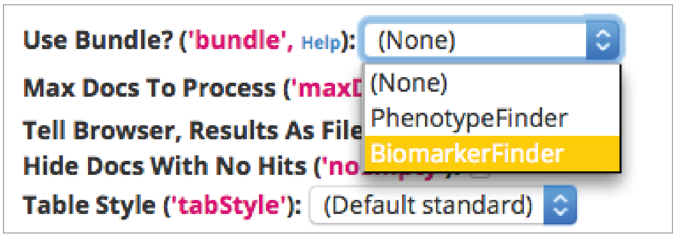
Having established that TERMite can identify all of the genes that are known markers in this corpus, we wanted to evaluate which were identified in phrases that matched the known and potential biomarker methodologies ("...gene is a good marker for disease..."). As can be seen from the data in **ROW 3** and **ROW 4** above, different linguistic patterns returned different biomarkers. The strongest evidence has been found for BRCA1, BRCA2, ERBB2, TOX3, MUC1 and PIP.

Table 1 shows the 4 gene biomarker candidates (NCOA3, PALB2, RAD54L and DCD) were not identified by methodologies 1a and 1b. We wanted to understand whether this was because the patterns of methodologies 1a and 1b missed these relations, or that the text was too complex to identify them. In all cases, we found no instance where the gene and indication were mentioned within the same sentence.

The abstracts do not provide any simple linguistic phrase that could be identified. This is a critical finding, highlighting the power of the approach used, combining highprecision and high-recall queries in one overarching search.
In conjunction with general cooccurrence, use of the MARKER/MARKERPLUS patterns give an excellent balance of precision and recall.

Simplified searching in TERMite/TExpress

Within TExpress, it is possible to create pattern bundles where a number of different search methodologies can be combined onto a single query. We have collated the three queries described above into a "Biomarker Finder" bundle. On the main interface, users can now run the analysis above through a single click without any need to review the details of all of the queries we have optimised.



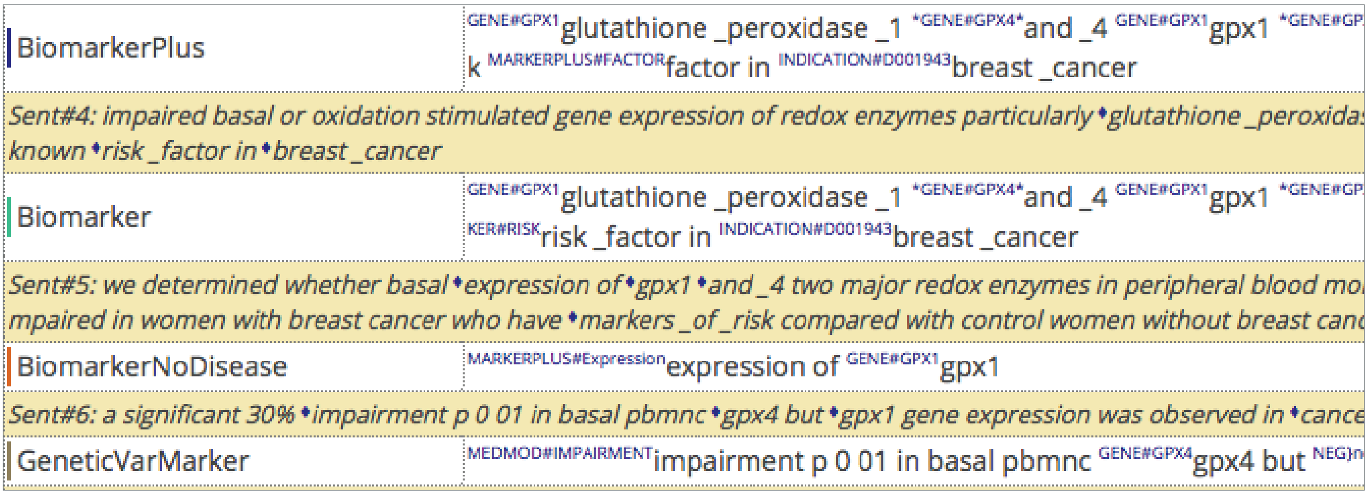
Use Bundle? ('bundle', Help): (None)

Max Docs To Process ('maxD

Tell Browser, Results As File

Hide Docs With No Hits ('no

Table Style ('tabStyle'): (Default standard)



BiomarkerPlus	GENE#GPX1 glutathione _peroxidase _1 *GENE#GPX4* and _4 GENE#GPX1 gpx1 *GENE#GP k MARKERPLUS#FACTOR factor in INDICATION#D001943 breast _cancer
Sent#4: impaired basal or oxidation stimulated gene expression of redox enzymes particularly •glutathione _peroxida known •risk _factor in •breast _cancer	
Biomarker	GENE#GPX1 glutathione _peroxidase _1 *GENE#GPX4* and _4 GENE#GPX1 gpx1 *GENE#GP KER#RISK risk _factor in INDICATION#D001943 breast _cancer
Sent#5: we determined whether basal •expression of •gpx1 •and _4 two major redox enzymes in peripheral blood mo mpaired in women with breast cancer who have •markers _of _risk compared with control women without breast canc	
BiomarkerNoDisease	MARKERPLUS#Expression expression of GENE#GPX1 gpx1
Sent#6: a significant 30% •impairment p 0 01 in basal pbmnc •gpx4 but •gpx1 gene expression was observed in •cance	
GeneticVarMarker	MEDMOD#IMPAIRMENT impairment p 0 01 in basal pbmnc GENE#GPX4 gpx4 but NEG}n

Figure 2: *In results users can see which methodology identified each gene-marker-disease combination*

Conclusion

SciBite's semantic software is designed to address many different topics across lifescience research. In this white paper we have discussed how we can enhance current published methodologies to scan biomedical literature and identify/extract reference to new and existing biomarkers for a given indication. If you would like to run a similar project to validate the performance of our capabilities, please get in touch.

Acknowledgements

SciBite would like to thank Ian Harrow from Ian Harrow Consulting for performing the data analysis described in this document.

SciBite
an ELSEVIER company

SciBite’s data-first, semantic analytics software is for those who want to innovate and get more from their data. At SciBite we believe data fuels discovery and we are leading the way with our pioneering infrastructure that combines the latest in machine learning with an ontology-led approach to unlock the value of scientific content. Supporting the world’s leading scientific organisations with use-cases from discovery through to development, SciBite’s suite of fast, flexible, deployable API technologies empower our customers, making it a critical component in scientific, data-led strategies. Contact us to find out how we can help you get more from your data.

To learn how SciBite can unlock the value of your data, speak to one of our experts today or email us at contact@scibite.com

Head Office:
SciBite Limited
BioData Innovation Centre
Wellcome Genome Campus
Hinxton, Cambridge CB10 1DR
United Kingdom

www.scibite.com
contact@scibite.com
LinkedIn: SciBite
Twitter: @SciBite
+44 (0)1223 786 129