

KLEIO is a knowledge-enriched information retrieval (IR) system developed at the UK National Centre for Text Mining (NaCTeM). The system offers textual and metadata searches across MEDLINE® and provides enhanced searching functionality by leveraging text mining technologies.

KLEIO draws upon a number of core technologies from the NaCTeM text mining tool kit to enhance automated detection and mark-up of biologically important terms appearing in text, such as gene/protein names. One of these tools is AcroMine which disambiguates acronyms based upon the context in which they appear. This functionality plays a key role in searching large document collections by allowing users to expand their queries and to include synonymous acronyms without losing the specificity of the original query.

The rich variety of term variants is a stumbling block for information retrieval as many forms have to be recognised, indexed, linked and mapped from text to existing databases. Typically, most of the currently available information retrieval systems (PubMed) fail to deal with the problems of term ambiguity and variability. For example, the term

$\{2-\}(\{3,4\}\text{-dihydroxy}\{\text{phenyl}\}\{\text{benzene}\})\}\{-\}\{\text{acetate}\}\{\text{acetic acid}\}$

can be expressed as

- 2-(3,4-dihydroxyphenyl)acetic acid, or
- 3,4-Dihydroxyphenyl acetate, or
- 3,4-Dihydroxybenzeneacetate

KLEIO addresses this problem by using our text mining technology for reducing the diversity of term variation. The conceptual approach to IR realised by KLEIO brings novel and original functionality to meet the growing interest in the biosciences looking for solutions to literature mining.

Without semantic annotation

- False Positive results occur due to similarity with non-protein entities
- False Negative results occur as search ignores synonym forms
- This results in poor accuracy and more than 60,000 results



With semantic annotation

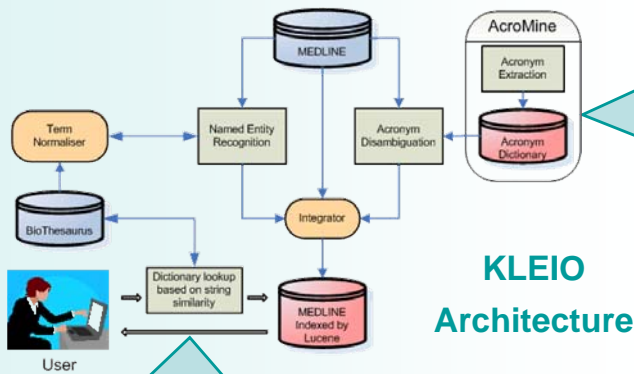
- Provides a more focused query
- Only documents with an annotated protein entity are returned
- Allows better integration with external protein databases and resources
- Fewer documents (reduced to 237)



Acronym recognition and disambiguation

- Recognizes acronyms and their long forms from the whole of MEDLINE®
- Disambiguates isolated acronyms using their context
- Maps acronyms into the corresponding long forms

A web demonstrator of this tool is available at: <http://www.nactem.ac.uk/software/acromine/>



KLEIO Architecture

Dictionary Lookup

Enter a gene/protein name: Submit Query

2 Human (Homo) c. cat c. Tact c. Drosophila (D) c. Cat

| Accession ID | Name | String Similarity |
|---------------------------------------|------|-------------------|
| BL.2 | | 0.99 |
| BL.2 mRNA | | 0.99 |
| BL.2 | | 0.99 |
| PTERLEKIN2 | | 0.79 |
| interleukin 2 | | 0.72 |
| interleukin 2 precursor | | 0.72 |
| interleukin 2 precursor | | 0.73 |
| interleukin 2 precursor | | 0.73 |
| interleukin 2 precursor | | 0.73 |
| human interleukin 2 | | 0.64 |
| interleukin 2 | | 0.63 |
| IFLA2protein | | 0.60 |
| IFLA2protein | | 0.61 |
| IFLA2PROTEIN | | 0.22 |
| ILP.2 | | 0.60 |
| Inhibitor of apoptosis-like protein 2 | | 0.50 |
| IAP-like protein 2 | | 0.53 |
| ILP.2 | | 0.60 |
| Interleukin-like transcript 2 | | 0.33 |
| IMRN/NOGLOBULIN/LK2 TRANSCRIPT 2 | | 0.24 |
| INTERLN_ALPHA.2 | | 0.87 |

Dictionary Lookup

An advanced dictionary service to:

- find synonyms and variants to assist in expanding a query across all relevant forms
- connect these forms with standardised identifiers (accession numbers)
- incorporate additional information from species specific databases

A web demonstrator of this tool is available at: <http://text0.mib.man.ac.uk/software/mlidc/>

References

- [1] <http://www.nactem.ac.uk/>
- [2] Ananiadou, S., Kell, D.B. and Tsujii, J. (2006) Text Mining and its potential applications in systems biology in Trends in Biotechnology, 12, 571-579
- [3] Ananiadou, S. & McNaught, J. (Eds) (2006) Text Mining for Biology and Biomedicine, Artech House Books.
- [4] Okazaki, N. and Ananiadou, S. (2006) Building an abbreviation dictionary using a term recognition approach, Bioinformatics, 22(24), 3089-3095.
- [5] Tsuruoka, Y., McNaught, J., Tsujii, J. and Ananiadou, S. (2007) Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. Bioinformatics, 23(20), 2768-2774.
- [6] Okanohara, D., Miyao, Y., Tsuruoka, Y., and Tsujii, J. (2006) Improving the Scalability of Sem'i-Markov Conditional Random Fields for Named Entity Recognition. Proceedings of Coling/ACL 2006, Sydney, Australia.
- [7] <http://lucene.apache.org/java/docs/>

For more information please visit:
<http://www.nactem.ac.uk/software/kleio/>
or contact
nactem@manchester.ac.uk