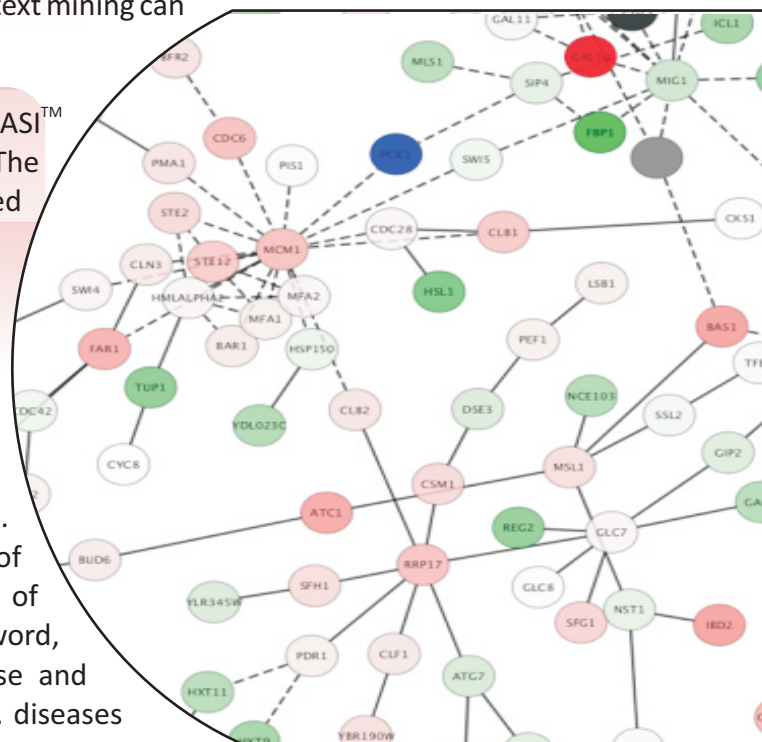


## Life Science Data Mining Solutions

Most of the publicly available biomedical data is in the form of abstracts and is semi-structured i.e. neither structured nor unstructured. There are structured fields like authors, references, keywords, title, date and also some unstructured fields like abstract, text or concepts. Also an important problem with the biomedical data is that a single term is linked to structures, sub-structures, Ids or pathways. It is also difficult to look for chemical names, especially IUPAC names, in plain text. Moreover, the biomedical literature is a complex set of information. It makes use of heavy domain specific terminologies. And extraction of useful information while maintaining all the links and relevance is quite a challenge. But the field of text mining can be used to solve such problems

We introduce LitSpec<sup>TM</sup>, a literature mining tool of RASIS<sup>TM</sup> suite which can be used to solve such problems. The unstructured cumbersome information can be converted into knowledge by LitSpec<sup>TM</sup>.

LitSpec<sup>TM</sup> uses the Machine learning-based approach also known as the statistical approach to parse the data. LitSpec<sup>TM</sup> uses dictionaries containing labeled and structured data. They can be used to extract biological keywords from text. LitSpec<sup>TM</sup> is powered by exhaustive dictionaries which are manually curated for better results. LitSpec<sup>TM</sup> also enables the users to view the classification of text into different categories by coloring the background of the related text words with same color. Given a keyword, LitSpec<sup>TM</sup> can harvest data from a pre-defined database and classify individual words into proteins, genes, chemicals, diseases and organisms.



### Deliverables

- A detailed delivery report for Life Science Data Mining service consists of:
- A chemical structure file (in case of chemical structures).
- A text, excel file for protein, gene and other biological terms.
- A network showing the interconnection between the terms found in the dataset.

### Required Input Data

- Journals, Books
- Patents
- Reports from Labs
- **Searching public domain databases (PubMed) for retrieving molecules, gene, protein, disease and organisms. Any changes can be done in the sentence structure**