

I2E Case Studies for Life Science

Text Search & Mining for Life Science

Case Study I: Reconstruction of a Biological Pathway from Medline Abstracts

Background

The mitogen-activated protein kinase (MAPK) pathways transduce a large variety of external signals, leading to a wide range of cellular responses. One of the major mammalian MAPK pathways, known as the MAPK/ERK pathway, is particularly associated with growth and differentiation.

Challenge

In this case study, scientists at Biowisdom wanted to validate the effectiveness of an Ontology Based Interactive Information Extraction (OBIIIE) system in finding known relationships in the pathway purely from the literature. Once validated, the system could then be used to mine for unknown relationships.

Solution

Linguamatics' I2E software was used in conjunction with Biowisdom's protein and relational ontologies to identify phrases that reconstruct the specific nature of protein-protein relationships in the MAPK pathway.

Benefit

The results obtained by the I2E system showed its value for precise and fast extraction of meaningful phrases from MEDLINE abstracts. Minimal effort (half a person day) set up queries that gave precise and accurate extraction of meaningful relationships from free text. Synonyms from the ontologies gave recall and I2Es linguistic queries gave precision.

Evaluation was focussed on RAF protein kinase and MAP kinase kinase, where a baseline precision of 12% (standard search for words in document) was increased to 73% using an I2E search for linguistically-motivated relations between ontology terms.

In this study I2E provided a straightforward way for non-linguists to leverage domain knowledge while performing linguistically-sophisticated queries. I2E searches can be run interactively in real time allowing users to rapidly prototype and develop queries which return the desired level of accuracy.

I2E, the Interactive Information Extraction software from Linguamatics, uses advanced natural language processing techniques that exploit the information contained in documents to produce content-rich output.

With I2E you can extract information from free text fast, with powerful interactive searches for specific information (e.g. what inhibits MEK?)

Case Study 2: Mining Nuclear Receptor Cofactors

Background

Nuclear receptors (NRs) are ligand-dependent transcription factors that typically recruit protein complexes (cofactors) to enhance or repress transcription of target genes. It is believed that several phenomena such as level of transactivation and tissue specificity of NRs depend heavily on the specific recruited cofactors. As NRs are important drug targets (18 of the 48 known human NRs are targets for registered drugs) the literature on these proteins is rapidly increasing.

Challenge

Researchers at AstraZeneca wanted to generate comprehensive and annotated lists of cofactors for several Nuclear Receptors (NRs): Liver X Receptor (LXR) a and b, and Androgen Receptor (AR).

Solution

AstraZeneca took Linguamatics' I2E software and plugged in a Biowisdom protein ontology and a modified relationship ontology. These were used to mine the MEDLINE and EMBASE literature databases for NR cofactors that were then compared to a "standard set" based on published and in-house lists.

Benefit

AstraZeneca achieved at least an order of magnitude speed-up over existing best practice, with no loss in quality or accuracy of the data. Furthermore, I2E discovered new cofactors not in the standard set, which were subsequently added to the standard set.

Typically, in a 1 person-day, 100 abstracts can be analyzed for cofactors. In this study, 2 person-hours with I2E extracted cofactors with a recall of 91% from around 8000 abstracts, rivalling hand analysis. This included 9 new and valid cofactors of AR not in the standard set.

Reusing the same query structures (i.e. no additional development time) for LXRs gave a recall of 90%, again rivalling hand analysis. These queries can now be reused again on much larger document sets. Because information in large collections is liable to be expressed in multiple ways, very high precision queries can achieve recall comparable to hand analysis, as in this study.

The case studies described show how I2E has been used to capture valuable information from the Life Sciences literature, saving time, and increasing productivity.