

GENIES

A Natural Language Processing Technology for the Extraction of Molecular Pathways from Complete Journal Articles

What Is GENIES?

GENIES is the GENomics Information Extraction System developed at the Columbia University Department of Biomedical Informatics by Carol Friedman and Andrey Rzhetsky. It extracts and structures information about cellular pathways from the biological literature in accordance with a knowledge model that was developed earlier.

GENIES Unique Advantage

GENIES is expected to be as accurate as scientists in its ability to abstract an exhaustive range of cellular pathways from complete narrative journal articles. It will deliver value with unsurpassed accuracy, breadth, and depth of understanding. Indeed, GENIES is a modified version of an existing medical natural language processing (NLP) system known as MedLEE (MEDical Language Extraction and Encoding). MedLEE is the only NLP system demonstrated to be as accurate as physicians in understanding narrative medical records. It is recognized as the best NLP for medical narratives in many medical subspecialties, and is currently used daily in the New York Presbyterian Hospital for patient care. GENIES inherits the robust and accurate functionalities of MedLEE, derived from twenty years of research and development.

Summary

GENIES

- extracts and structures information about cellular pathways,
- is based on seasoned natural processing technologies that have repeatedly shown exceptional accuracy, breadth and depth, comparable to human experts,
- has rich and flexible input and output, and is XML compliant,
- offers rapid and predictable product development and delivery based on a modular architecture that separates the thoroughly tested PROLOG processor from the development of new knowledge sources.

Development Stage

GENIES' architecture and programming is completed. The knowledge engineering in the lexicon component is limited to signal transduction pathways (a subset of all cellular pathways) and is currently under development. The precision of GENIES in a published preliminary evaluation is 96%.

Applications

A crucial component of innovative knowledge management tools, GENIES provides timely access to relevant biological data and, thus, enables life science researchers to make better decisions faster.

It serves as a core pipeline to deliver information to knowledge management tools, cellular model systems, virtual cellular experimental systems, virtual clinical and pre-clinical trials. It also can associate knowledge derived from the literature with information derived from data mining techniques applied to high throughput technology data.

Intellectual Property Position

A patent application has been filed.

Technology Position

GENIES technology is available for licensing.

Technical Contact

Carol Friedman, Ph.D., Professor, Department of Biomedical Informatics, Columbia University,
Andrey Rzhetsky, Ph.D., Associate Professor, Department of Biomedical Informatics and Columbia
Genome Center, Columbia University

For more information

Vincent Tomaselli, Deputy Director; Center for Advanced Technology, *voice*: 212.305.2944; *e-mail*:
tomaselli@dmi.columbia.edu

CAT Technology Overview

The Center for Advanced Technology (CAT) in Information Management at Columbia University is a joint effort of the Department of Medical Informatics (Columbia Presbyterian Medical Center), Computer Science Department (School of Engineering, Columbia University), and Columbia Genome Center.

Medical Informatics deals with the storage, retrieval, sharing, and optimal use of biomedical information, data, and knowledge for problem solving and decision-making. It touches on all basic and applied fields in biomedical science, and is closely tied to modern information technologies, notably in the areas of computing and communications.

Researchers in the *Computer Science Department* study theoretical and experimental aspects of many areas of information management and technology - foundations in mathematics, optimization, hardware design, software design, networks, user interfaces, databases, communications, and artificial intelligence. In particular, the department has concentrated expertise in digital library technology, digital government systems, and novel visual and graphical interfaces for information management.

The *Columbia Genome Center* is a consortium of University scientists working on gene discovery and technology development related to the human genome. Integrated genomic mapping and sequencing is used to facilitate gene discovery and gene therapy strategies. The aim is to encourage the rapid clinical application of new developments by providing all areas of expertise and resources that are necessary for the development of novel diagnostics and therapeutics.

CAT Mission

The goal of the Centers for Advanced Technology program is to support cutting-edge research at major New York State research institutions, and to make the resulting technology available to industry for commercialization. CATs work with industry partners in several ways to achieve this goal. Inquiries are welcomed.

The Center for Advanced Information Management is sponsored by the New York State Office of Science, Technology, and Academic Research

