

GeneWays

High Throughput Extraction, Compilation, and Analysis of Molecular Pathways from Complete Journal Articles

What Is GeneWays?

[GeneWays](#) software compiles and analyzes molecular pathways based on automated extraction of massive amounts of information from research literature.

Advantages

[GeneWays](#) can abstract an exhaustive range of cellular pathways from all narrative journal articles available electronically, delivering value with unsurpassed accuracy, breadth, and depth.

Development Stage

Four core modules of [GeneWays](#) have been developed, integrated, and evaluated:

- Tagger¹ identifies and tags genes and proteins,
- GENIES^{2,3}, a natural language processing system that identifies and extracts biomolecular relationships,
- Knowledge Model⁴ represents biomolecular information, and
- Visualization System⁵, a graphical tool for the presentation of pathways.

Two other options, a disambiguation⁶, and a statistical data integration⁷ system, have also been developed and evaluated and are currently being integrated to [GeneWays](#).

Summary

[GeneWays](#) -

- Extracts, disambiguates, curates, structures, edits, visualizes and analyzes information about molecular pathways;
- Is based on seasoned natural language processing and knowledge-based modeling technologies that have repeatedly shown exceptional accuracy, breadth and depth, comparable to human experts;
- Has rich and flexible input and output; and
- Has rapid and predictable product development and delivery based on a modular architecture that separates the thoroughly tested PROLOG³ processor from the development of new knowledge sources or modules.

Another module in progress will perform automated maintenance of a knowledge base that contains comprehensive information about molecular pathways, diseases associated with the pathways, and the drugs that affect them.

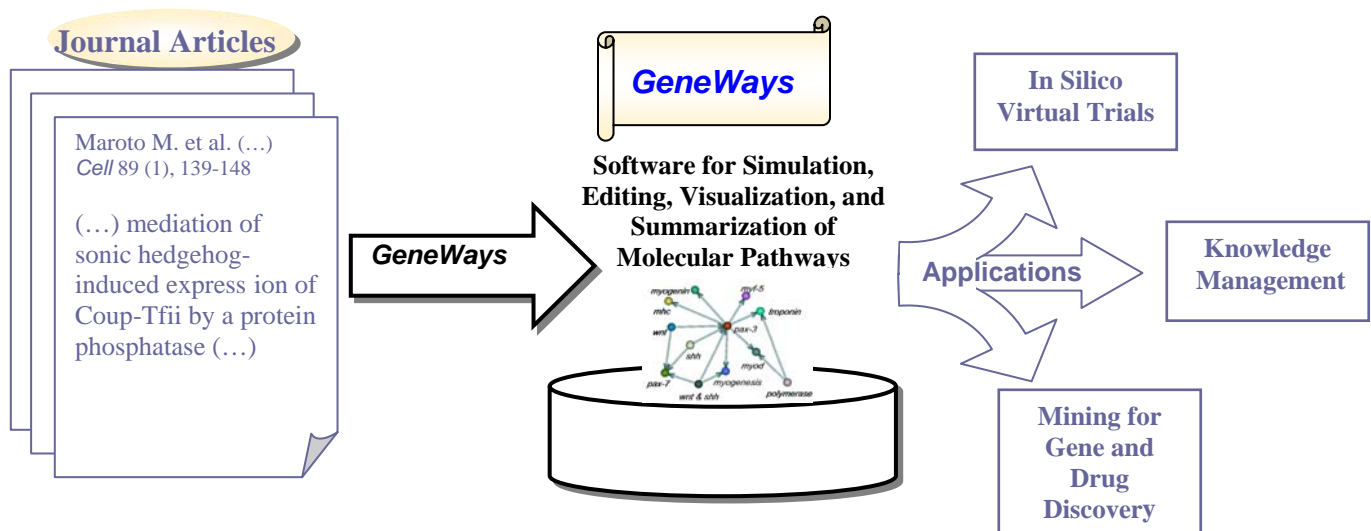
The knowledge engineering in the lexicon component is presently geared to signal transduction pathways (a subset of all cellular and biochemical pathways). In a published preliminary evaluation, the ability of [GeneWays](#) to extract information was reported to be 96%^{1,2}.

Applications

[GeneWays](#) provides a crucial component of innovative knowledge management tools for timely access to relevant biological data. It thus enables life science researchers to make better decisions faster, and ultimately leads to faster drug research and development. For example, it can be used as a core pipeline to deliver information to visualization and summarization tools, knowledge management tools, cellular model systems, virtual cellular experiment systems, and virtual clinical and pre-clinical trials. It can also be used to associate knowledge derived from the literature with information derived from data mining techniques applied to high throughput technology data.

Schematic Representation of GeneWays

The Ability to Increase Information and Business Value from Published Literature



References

- (1) Krauthammer M., Rzhetsky A., Morozov P., and Friedman C. (2000), *Using BLAST For Identifying Gene And Protein Names In Journal Articles*. *Gene*, **259**, 245-252. <http://genome6.cpmc.columbia.edu/andrey/MichaelGene.pdf>
- (2) Friedman C, Kra P., Krauthammer M, Yu H, Rzhetsky A, (2001), *GENIES: A Natural-Language Processing System for the Extraction of Molecular Pathways from Journal Articles*. (to appear in *Bioinformatics*).
- (3) Friedman C A Broad Coverage Natural Processing System. Proc AMIA Symp. 2000;270-4.
- (4) Rzhetsky A., Koike T., Kalachikov S., Gomez S. M., Krauthammer M., Kaplan S. H., Kra P., Russo J. J., and Friedman C. (2000), *A Knowledge Model For Analysis And Simulation Of Regulatory Networks*. *Bioinformatics*, **16**, 1120-1128. <http://genome6.cpmc.columbia.edu/andrey/OntologyBioinformatics.pdf>
- (5) Koike, T., and A. Rzhetsky. 2000. *A Graphic Editor For Analyzing Signal-Transduction Pathways*, *Gene* **259**: 235-244 <http://genome6.cpmc.columbia.edu/andrey/TomohiroGene.pdf>
- (6) Hatzivassiloglou V., Duboue P. A., and Rzhetsky A. (2001) *Disambiguating Proteins, Genes, And RNA In Text: A Machine Learning Approach*, (accepted by *Bioinformatics*).
- (7) Gomez S. M., Lo S. H., and Rzhetsky A. (2001) *Probabilistic Prediction Of Unknown Metabolic And Signal-Transduction Networks*. (accepted by *Genetics*).
- (8) Friedman, C., Kra, P. & Rzhetsky, A. (2002) *Two biomedical sublanguages: a description based on the theories of Zellig Harris*. *J Biomed Inform*, **35**, 222-235.
- (9) Krauthammer, M., Kra, P., Iossifov, I., Gomez, S. M., Hripcsak, G., Hatzivassiloglou, V., Friedman, C., and Rzhetsky, A. (2002). *Of truth and pathways: chasing bits of information through myriads of articles*. *Bioinformatics* **18 Suppl 1**, S249-S257. (ISMB 2002).

Intellectual Property Position

A patent application has been filed.

Technology Position

GENIES technology is available for licensing.

Technical Contacts

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

Carol Friedman, Ph.D.; Professor, Department of Biomedical Informatics, Columbia University

For more information

Vincent Tomaselli, Deputy Director; Center for Advanced Information Management, voice: 212.305.2944;

e-mail: tomaselli@cat.columbia.edu