Imitating manual curation of text-mined facts in biomedicine

Raul Rodriguez-Esteban [1,2], **Ivan Iossifov** [2,3], **Andrey Rzhetsky** [2,3,4]

[1]*Department of Electrical Engineering,* [2]*Center for Computational Biology and Bioinformatics, Joint Centers for Systems Biology,* [3]*Department of Biomedical Informatics,* [4]*Judith P. Sulzberger MD Columbia Genome Center and Department of Biology, Columbia University, New York, NY 10032, U.S.A.*

Text-mining algorithms make mistakes in extracting facts from the natural-language texts. In biomedical applications, which rely on use of text-mined data, it is critical to assess the extraction quality (the probability that the message is correctly extracted) of individual facts. Using a large set of almost 100,000 manually produced evaluations (most facts were independently reviewed more than once producing independent evaluations), we implemented and tested a collection of algorithms that mimic human evaluation of facts provided by an automated information-extraction system [1]. The algorithms that were used include several Bayesian classifiers, SVMs, Neural Networks and Maximum Entropy methods. The performance of our best automated classifiers, a second-order Maximum Entropy classifier, closely approached that of our human evaluators (ROC score close to 0.95). Were we to use a larger number of human experts to evaluate any given sentence, we could implement an artificial-intelligence curator that would perform the classification job at least as accurately as an average individual human evaluator. Hence we present a system that automatically curates the interactions that are extracted from the biomedical literature. This system is useful for enhancing the quality of information gathered by text-mining techniques. Below, we illustrate our analysis by visualizing the predicted accuracy of the text-mined relations involving cocaine (see Figure 1).

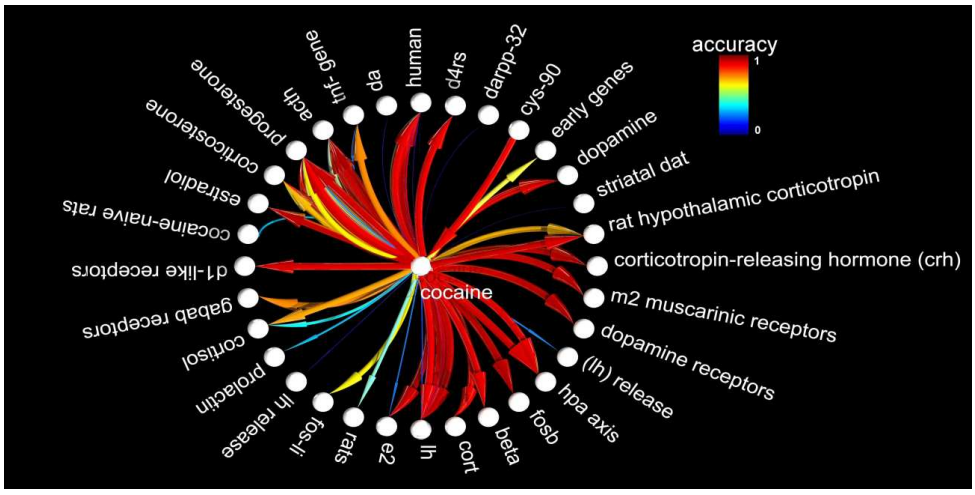

Figure 1: *Cocaine*: the predicted accuracy of individual text-mined facts involving semantic relation *stimulate* in *Geneways 6.0* is shown both in color and in width of the corresponding arc.

[1] Rzhetsky A, et al. (2004) GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. J Biomed Inform 37:43-53.