MolecRank: A Network Analysis Algorithm for Ranking Therapeutic Molecules in Biological Networks

Ahmed Abdeen Hamed · Agata Leszczynska · Mark Schreiber

Abstract Pharmaceutical scientists often search databases of therapeutic molecules (i.e. drugs) to answer a set of drug discovery queries. In this paper, we present a novel algorithm called MolecRank, which is a molecule ranking mechanism. The algorithm traverses a network of biological features (e.g., therapeutic molecules, genes, protein, chemical compounds, cell-type, RNA, disease name, etc.) extracted from publications (i.e., article abstracts). Starting with the PubMed web portal, we searched for two specific keywords to guarantee relevant abstracts: "Merck", and "MK" for the MK Number that encodes a Merck molecule. The PubMed query resulted in 792 publicly available MEDLINE abstracts.

From the biological features extracted, we constructed a network: its nodes are (genes, diseases, and chemical compounds), while its links are co-occurrence incidences between a molecule and the feature within the same abstract. That is: the nodes are connected based on their mentions in the same article. If two nodes (e.g., a molecule and a disease) are mentioned in the same abstract, they are linked. The network is stored in a graph database (a triplestore) to make it accessible for querying. When a query is issued against the triplestore database, the algorithm performs a post-processing ranking step. Its purpose is to returns the most relevant and most specific molecules first. For any given query, the result result shows interesting and produces unique rankings for the molecules in the network. Such ranking can not be achieved using a single network centrality measure (e.g., degree, eccentricity, closeness, betweenness, or PageRank), which is known to be the gold-standard in such cases.

A. Hamed
Merck & Co., Inc., Boston, MA, USA
E-mail: ahmed.hamed1@merck.com

A. Leszczynska
MSD, Prague, Czech Republic

M. Schreiber
Merck & Co., Inc., Boston, MA, USA