Michael Krauthammer and Thaibinh Luong

Term Mapping Using Matrix Operations

We believe that gene name identification is a modular process involving term recognition, classification and mapping. This work's focus is on gene name mapping, and we assume that names are already recognized and classified. We use a combination of two methods to map recognized entities to their appropriate gene identifiers (Entrez GeneIDs): the Trigram Method, and the Network Method. Both methods require preprocessing, using resources from Entrez Gene, to construct a set of method-specific matrices. We first address lexical variation by transforming gene names into their unique "trigrams" (groups of three alphanumeric characters), and perform trigram matching against the preprocessed gene dictionary. For ambiguous gene names, we additionally perform a contextual analysis of the abstract that contains the recognized entity. We have formalized our method as a sequence of matrix manipulations, allowing for a fast and coherent implementation of the algorithm. In this talk, we also show how gene name identification, and text mining in general, can play a critical role in translational medicine. We demonstrate how term identification is useful for establishing a biobibliometric distance between genes and psychiatric disorders.

Dagstuhl Seminar Proceedings 08131
Ontologies and Text Mining for Life Sciences : Current Status and Future Perspectives
http://drops.dagstuhl.de/opus/volltexte/2008/1512