Poster K-3

Disambiguation of large gene/protein terminologies by automatic literature profiling

Authors:
Roney S Coimbra (*GlaxoSmithKline - Cheminformatics*)
Dana E. Vanderwall (*GlaxoSmithKline - Cheminformatics*)

Short Abstract: We present a method to disambiguation of gene/protein terminologies. Aliases of a given gene produce collections of Medline abstracts which share a common vocabulary. This gene/protein specific vocabulary defines a unique fingerprint that can be used to disclose ambiguous aliases.

Long Abstract:
In recent years, text mining of scientific literature has proved to be a powerful tool to define networks of protein-protein interactions. These interactions can be inferred based on the co-occurrence frequency of gene/protein names in document sets, or, more precisely, by parsing the syntactic/semantic status of two protein names in phrases where they occur linked by verbs defining the type/mechanism of interaction between them. Both approaches require high quality ontology of interaction types/mechanisms, and a disambiguated terminology of gene/protein aliases. We present herein an original method of disambiguation of large gene/protein terminologies for which manual curation would be hardly achievable.
Aliases of a given gene produce collections of Medline abstracts sharing a common vocabulary. This is true even when these aliases are not associated with the same subset of documents. This gene/protein specific vocabulary defines a unique fingerprint that can be used to disclose ambiguous aliases. We define a case as a gene/protein alias, and a group as the whole set of aliases for a given gene/protein. Each case in each group was used to query Medline for recent abstracts, separately. These collections of abstracts were analysed for vocabulary frequencies using a previously published algorithm [1]. The retrieved vocabulary was filtered to eliminate too broadly occurring terms which appear in abstracts of most groups and may not help defining group identity. For each group (gene), cases (aliases) were clustered by hierarchical clustering according to the vocabulary in their respective Medline abstracts. Outliers with low correlation with the main cluster were labelled “ambiguous”.

[1]Chaussabel D & Sher A, Mining microarray expression data by literature profiling. Genome Biol. 2002 Sep 13;3(10):RESEARCH0055.