Mapping proteins to disease terminologies: from UniProt to MeSH

MOTTAZ, Anais, et al.

Abstract

BACKGROUND: Although the UniProt KnowledgeBase is not a medical-oriented database, it contains information on more than 2,000 human proteins involved in pathologies. However, these annotations are not standardized, which impairs the interoperability between biological and clinical resources. In order to make these data easily accessible to clinical researchers, we have developed a procedure to link diseases described in the UniProtKB/Swiss-Prot entries to the MeSH disease terminology. RESULTS: We mapped disease names extracted either from the UniProtKB/Swiss-Prot entry comment lines or from the corresponding OMIM entry to the MeSH. Different methods were assessed on a benchmark set of 200 disease names manually mapped to MeSH terms. The performance of the retained procedure in term of precision and recall was 86% and 64% respectively. Using the same procedure, more than 3,000 disease names in Swiss-Prot were mapped to MeSH with comparable efficiency. CONCLUSIONS: This study is a first attempt to link proteins in UniProtKB to the medical resources. The indexing we provided will help clinicians and researchers navigate from [...]

Reference


DOI : 10.1186/1471-2105-9-S5-S3
PMID : 18460185
Regular Expressions used to extract the disease names from the Swiss-Prot disease comment lines

<table>
<thead>
<tr>
<th>(1) Starter expressions</th>
<th>(2) Specific stop words</th>
<th>(3) Termination term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cause(s) of /a involved in</td>
<td>susceptibility to development of genetic predisposition for developing pathogenesis of subset of various types of some form of increased risk of</td>
<td>also known as but which an due to in condition(s) such as . [MIM:</td>
</tr>
<tr>
<td>(can) contribute(s) to associated/association with correlated with responsible for contributor to result(s)/resulting in lead(s) to induce(s) defective in individual(s) with patient(s) with/suffering from reduce(s) influence(s) deleted in down-regulated in found in implicated in predispose(s) to favor antigen of antigen for thought to be an role in could impart mediate(s) candidate (gene)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(1) Expressions used to extract the part of the string containing the disease name. (2) Terms removed from the string extracted. (3) Expressions indicating the end of the disease name.