Data mining for drug discovery, exploring the universes of chem- and bioinformatics

Modest von Korff, Thomas Sander
Research Information Management
Actelion Pharmaceuticals Ltd.
Gewerbestrasse 16
4123 Allschwil
Switzerland

modest.korff@actelion.com

To find new chemical entities which cure or palliate a disease is the most important task in pharmaceutical research. Data mining in pharmaceutical research has to support this task and here an overview is given how data mining may contribute to drug discovery. During the last fourteen years our research group developed a multitude of data mining tools for drug discovery. This includes the field of chemistry, where many challenges are still waiting, as well as biology where the tasks of data mining become a snowballing list. Data mining in chemistry is working on chemical structures and their properties. The total number of possible chemical structures, the chemical space, is almost endless. Data mining has to blaze a trail in the chemical space to detect new bioactive chemical entities. Relating chemical structures with bioactivity opens a new universe the bio-chem-space, which is characterized by the conformations of the chemical structures and their relations to target proteins. Target proteins are large molecules which act as switches for cells. A bioactive molecule triggers the switch-function at a protein. This may initiate a change in the function of the cell were the protein is located. If the function of the cell is related to a disease the change in the behaviour of the cell may improve the symptoms of the disease. Mining for information on these complex relationships is the task of bioinformatics. This is here exemplified by bioactivity database mining, image analysis, and mining scientific medical literature. An efficient approach in mining scientific literature for mapping genes to diseases demonstrates the research work in our lab. Detecting non-obvious relations between genes and diseases opens new opportunities to find indications for new chemical entities. For this task we developed a new algorithm G2DPubMedMiner and challenged it by six third-party tools. For the challenge a new test dataset with pharmaceutical-relevant gene-disease associations was introduced. This test dataset relies on the triple association between an approved drug, a disease, and a gene. Approved drugs were chosen because they provide strong evidence for the triple association between the drug, a target protein, and the gene that encodes the target protein. A set of 39 drugs was selected to create the test dataset. G2DPubMedMiner relies on a straightforward algorithm. PubMed Central is queried with a gene name and its synonyms. Retrieved result records are filtered and the accepted records are indexed by disease-related MeSH terms. The MeSH terms found are ranked by their frequency of occurrence in the result records. To validate G2DPubMedMiner, the program was fed with the gene names from the test dataset GDtest. For each gene, a table that contained the disease MeSH terms and their frequency of occurrence was compiled. The frequency of occurrence determined the rank of the disease MeSH term. From this rank, the relative rank was calculated, and received a number between zero and one. This relative rank was used as figure of merit for the relevance of the gene-disease association. The relative ranks for all test records from the GDtest were calculated for the G2DPubMedMiner as well as for the following six third-party tools: DISEASES, DisGeNET, HuGENavigator, Ingenuity, MalaCards, and NextBio. Some details of the developed algorithm are presented and the result of the tool comparison is given in this contribution.