# PubClust: a tool for conceptual organization of PubMed abstracts

Fattore M.(1),Arrigo P.(1)

1) CNR ISMAC Via De Marini 6 16149 Genova

## Motivation

One of the major challenges in the post-genomic era is the improvemnt of the capability to recognize the disease related molecular targets. Even if the new experimental methods have greatly enhanced the analytical capability,the application of NLP (natural language processing) play a relevant role in the selection and extraction of relevant information. The published paper still constitute the main entry point to bridge togheter molcular biology, chemical and medical information. The text processing gives a great effort in the speed up the data warehousing. In this paper we present an interactive tool that allow to conceptually organize organize the Medline.

## Methods

The system combine an 'unsupervised' Machine Learning method with a statistical clustering approach. The analysis id performed on the Medline abrstracts, withouth the use of any external support information such as MESH terms or predefine ontologies. The document set is retrieved from PubMed DB,according the NCBI e-tools, and,consequently they are linguistically preprocessed. The ML module allows a preliminary classification of these documents. An iterative agglomerative procedure is appliec on the set of activated neurons in order to build a conceptual hierarchy withouth a predefined knowledge. The use of this approach allows to define the intirinsic term dependency.

## Results

This system allows to obtain a set of association rules directly extracted from the abstract text. The simple topic classification is not enough to evaluate the content homogeneity of a document set. The identification of semantically correlated document is still a complex and very hard task. The PubClust system allows the researcher to select specific conceptually homogeneous set of abstracts in order to reduce the time required for document screening.

**Contact email:** arrigo@ge.ismac.cnr.it
**URL:** http://biocomp.ge.ismac.cnr.it/