

Knowledge Discovery in Neuroinformatics

Bartłomiej Wilkowski

Technical University of Denmark, DTU Informatics

bw@imm.dtu.dk

Traditionally, the process of turning data into bio-medical knowledge has involved manual, meta-analysis of results reported in journals. Since the amount of the scientific data produced nowadays is increasing very dramatically, there is a need and room for tools which could enable faster data processing and its statistical analysis. Moreover, such a rapid growth of data and a resultant expansion of the medical databases has created a significant potential for creation of efficient data modelling and information retrieval tools.

The Center for Integrated Molecular Brain Imaging (CIMBI)¹ with which this PhD project is associated, has established a large database of behavioral, genetic and imaging data. The key challenge of the research to be carried out during the PhD-study is to develop the methods for integration of the CIMBI and related distributed databases including literature, biobanks and DTU's functional imaging database "Brede" in order to create an intelligent service for efficient information retrieval. Such a service is likely to become important not only for extracting information but also for an assistance in various aspects of research such as discovery of new facts, identification of previously undiscovered associations followed by proposal of new functions [2]. The employment of any kind of "informal" databases, containing unstructured data such as scientific publications, with the above-mentioned service requires development of the techniques for automatic retrieval of the relevant functional information.

In this PhD-study ontologies (e.g. UMLS ontology) and formal and statistical methods (e.g. SVD, NMF) are considered to be key tools for both database integration as well as the development of a process which is referred to as "knowledge discovery". Such a process can be understood as a pipeline of methods and techniques which include: text-mining of the scientific publications and further information retrieval (keyword extraction), automatic interpretation of findings, discovery of new relationships and even design of new experiments.

The study will depart from the in-house CIMBI and Brede biobanks and aim towards their further integration with key international databases such as UMLS, PubMed [1], and specialized neuroinformatics tools.

References

- [1] Andreas Doms and Michael Schroeder. GoPubMed: exploring PubMed with the Gene Ontology. *Nucl. Acids Res.*, 33(2):W783–786, 2005.
- [2] M. Krallinger and A. Valencia. Text-mining and information-retrieval services for molecular biology. *Genome Biol.*, 6(7), 2005.

¹CIMBI is a research center funded by the Lundbeck Foundation. The research in the Center focuses on the neural bases of personality dimensions that predispose individuals to affective and substance use disorders, with special emphasis on the serotonergic neurotransmitter system. Both PET and MRI are employed in studies of human subjects, and these are complemented with relevant studies using animal models. Advanced informatics techniques, new tracer compounds, and novel serotonergic challenge paradigms are also being developed within the center. The work venues span several institutions in Copenhagen as well as several collaborating laboratories in Europe and the U.S.