Keynote Address

Chemoinformatics: An Application Domain for Information Retrieval Techniques

Peter Willett
Department of Information Studies, University of Sheffield
Sheffield S10 2TN, U.K.
+44-114-2222633 / p.willett@sheffield.ac.uk

ABSTRACT

Chemoinformatics is the generic name for the techniques used to represent, store and process information about the two-dimensional (2D) and three-dimensional (3D) structures of chemical molecules [1, 2]. Chemoinformatics has attracted much recent prominence as a result of developments in the methods that are used to synthesize new molecules and then to test them for biological activity. These developments have resulted in a massive increase in the amounts of structural and biological information that is available to support discovery programmes in the pharmaceutical and agrochemical industries.

Chemoinformatics may appear to be far removed from information retrieval (IR), and there are indeed many significant differences, most notably in the use of graph representations to encode chemical molecules, rather than the strings that are used to encode text; however, there are also many similarities between the two fields, and this paper will exemplify some of these relationships. The most obvious area of similarity is in the principal types of database search that are carried out, with both application domains making extensive use of exact match, partial match and best match searching procedures: in the IR context these are known-item searching, Boolean searching and ranked-output searching; in the chemical context, these are structure searching, substructure searching and similarity searching. In IR, there is a natural distinction between an initial ranked-output search and one in which relevance feedback can be employed, where the keywords in the query statement are assigned weights based on their differential occurrences in known-relevant and known-nonrelevant documents. In the chemoinformatics technique called *substructural analysis*, substructural fragments are assigned weights based on their occurrence in molecules that do possess, and molecules that do not possess, some desired biological activity [3]. The analogy between relevance and biological activity has also resulted in the development of measures to quantify the effectiveness of chemical searching procedures that are based on the standard IR concepts of recall and precision [4].

Analogies such as these have provided the basis for some of the chemoinformatics research carried out in Sheffield. The starting point was the recognition that techniques applicable to documents represented by keywords might also be applicable to molecules represented by substructural fragments. This led directly to the introduction of similarity searching, something that is now a standard tool in chemoinformatics software systems; in particular, its use for *virtual screening*, i.e., the ranking of a database in order of decreasing probability of activity so as to maximize the cost-effectiveness of biological testing [5]. Measures of inter-molecular structural similarity also lie at the heart of systems for clustering chemical databases: just as IR has the Cluster Hypothesis (similar documents tend to be relevant to the same requests) as a basis for document clustering, so the Similar Property Principle (similar molecules tend to have similar properties) has led to clustering becoming a well-established tool for the organization of large chemical databases [6]. More recently, we have applied another IR technique, the use of data fusion to combine different rankings of a database, to chemoinformatics and again found that it is equally applicable in this new domain [7].

The many similarities between IR and chemoinformatics that have already been identified suggest that chemoinformatics is a domain of which IR researchers should be aware when considering the applicability of new techniques that they have developed.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering, Relevance feedback. H.3.7 [**Digital Libraries**]. J.2 [**Physical Sciences and Engineering**]: Chemistry

General Terms

Algorithms, Performance, Design, Experimentation.

Keywords: Chemistry, Chemoinformatics, Cluster Hypothesis, Data fusion, Molecules, Relevance feedback, Similarity searching, Substructure searching, Virtual screening

REFERENCES

1. Leach, A.R. and Gillet V.J. *An Introduction to Chemoinformatics*. Kluwer, Dordrecht, 2003.
2. Gasteiger, J. and Engel, T. *Chemoinformatics. A Textbook.* Wiley-VCH, Weinheim, 2003.
3. Ormerod, A., Willett, P. and Bawden, D. Comparison of fragment weighting schemes for substructural analysis. *Quant. Struct.-Activ. Relat.*, *8* (1989), 115-129.
4. Edgar, S.J., Holliday, J.D. and Willett, P. Effectiveness of retrieval in similarity searches of chemical databases: a review of performance measures. *J. Mol. Graph. Model.*, *18* (2000), 343-357.
5. Willett, P., Barnard, J.M. and Downs, G.M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, *38* (1998), 983-996.
6. Willett, P. *Similarity and Clustering in Chemical Information Systems*. Research Studies Press, Letchworth, 1987.
7. Ginn, C.M.R., Willett, P. and Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discov. Design*, *20* (2000), 1-16.