Visually Guiding and Controlling the search while Mining Chemical Structures

Max Pereira[1] and Vítor Santos Costa[2] and Rui Camacho[1] and Nuno A. Fonseca[3,2]

[1] LIAAD-INESC Porto LA & FEUP, Universidade do Porto,
Rua Dr Roberto Frias s/n, 4200-465 Porto, Portugal
[2] CRACS-INESC Porto LA, Universidade do Porto,
Rua do Campo Alegre 1021/1055, 4169-007 Porto, Portugal
[3] Instituto de Biologia Molecular e Celular (IBMC), Universidade do Porto
Rua do Campo Alegre 823, 4150-180 Porto, Portugal

Abstract. In this paper we present the work in progress on LogCHEM, an ILP based tool for discriminative interactive mining of chemical fragments. In particular, we describe the integration with a molecule visualisation software that allows the chemist to graphically control the search for interesting patterns in chemical fragments. Furthermore, we show how structured information, such as rings, functional groups like carboxyl, amine, methyl, ester, etc are integrated and exploited in LogCHEM.

Keywords: Inductive Logic Programming, drug design

1 Introduction

Structural activity prediction is one of the most important tasks in chemoinformatics. The goal is to predict a property of interest given structural data on a set of small compounds or drugs. This task can be seen as an instance of a more general task, *Structural Activity Regression* (SAR), where one aims at predicting activity of a compound under certain conditions, given structural data on the compound. Ideally, systems that address this task should not just be accurate; they should be able to identify an *interpretable* discriminative structure which describes the most discriminant structural elements with respect to some target.

LogCHEM leverages the flexibility of the Inductive Logic Programming (ILP)[1] learning paradigm while addressing the three main principles enunciated above. We demonstrated that LogCHEM can be used to mine effectively large chemoinformatics data sets, such as the DTP AIDS data set [2]. LogCHEM can input data from chemical representations, such as MDL's SDF file format, and display molecules and matching patterns using visualisation tools such as VMD [3]. The structure of LogCHEM is shown in Figure 1.

Ultimately, our goal is for LogCHEM to become a truly interactive system for drug discovery, *iLogCHEM*. In this work, we present a step forward in this direction. In *iLogCHEM*, we want to allow users to participate in the drug discovery process in a number of ways:

1. We propose the ability to incorporate user-provided abstractions, of interest to the chemoinformatics domain, that can be used to aid the discovery process. As a first experiment, we have allowed users to specify a common chemical structure, *aromatic rings*. The user has available in *iLogCHEM*, apart from the *aromatic rings*, functional groups such as *carboxyl*, *amine*, *ester*, *methyl*, *phenyl* etc
2. We propose an interactive refinement process where the user can interact with the proposed model, adapting it, evaluating it, and using it to guide (constrain) the search.

Next, we motivate the two main goals of *iLogCHEM*.

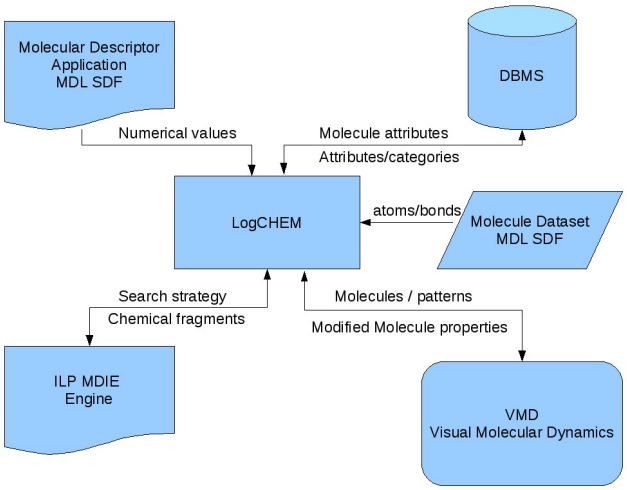

Fig. 1. LogCHEM system

2 Representing and Displaying Molecules

In order to fulfil our goals, the first problem that one has to address is how to describe molecules. Coordinate-based representations usually operate by generating features from a molecule's 3D-structure [4]. The number of features of interest can grow very quickly, hence the problem that these systems need to address is how to select the most interesting features and build a classifier from them. Coordinate-free representations can use atom pair descriptors or just the atom-bond structure of the molecule. In the latter case, finding a discriminative component quite often reduces to the problem of finding a Maximum Common Substructure (MCS).

LogCHEM follows the latter approach. LogCHEM uses a logic representation, where atoms and bonds are *facts* stored in a database. Although our representation is less compact than a specialised representation such as SMILES, used

in MOLFEA [5] and SMIREP [6], it offers a number of important advantages. First, it is possible to store information both on atoms and on their location: this is useful for interfacing with external tools. Second, LogCHEM can take advantage of the large number of search algorithms implemented in ILP. Third, given that we implement the basic operations efficiently, we can now take advantage of the flexibility of our framework to implement structured information.

LogCHEM was originally built as a three step pipeline. First, chemical data is filtered to our logical format. Second, we use the LogCHEM ILP learner to generate rules. Last, the rules and how they fit the examples are displayed using a tool such as `VMD` [3] to display the molecules and the matching substructures. Figure 2 shows an example pattern for the HIV data set. The pattern is shown as a wider atoms and bonds, and it includes a sulphur atom and part of an aromatic ring.

3 Macros

A first step forward stems from observing Figure 2: does the pattern include part of the ring because only part of the ring matters or, as it is more natural from the chemists point of view, should we believe that the whole ring should be in the pattern. Quite often discriminative miners will only include part of a ring because it is sufficient for classification purposes. But this may not be sufficient to validate the pattern.

The logical representation used in LogCHEM makes it natural to support *macro* structures, such as rings used in MoFa [7] in a straightforward fashion. The next example shows such a description:

```
macro(M,(atom(A1,c), bond(A1,A2,_),
 atom(A2,c), bond(A2,A3,_),
 atom(A3,c), bond(A3,A4,_),
 atom(A4,c), bond(A4,A5,_),
 atom(A4,c), bond(A4,A5,_),
 atom(A5,c), bond(A5,A6,_),
 atom(A6,c), bond(A6,A1,_))).
```

Initial results with LogCHEM show that using this macro results in similar accuracy, but *more easy to interpret rules*. In *iLogCHEM* we are following two directions: a library of preexisting common patterns, that will be immediately available for discovery, and the ability to define a new pattern *graphically* and then translate it to the LogCHEM internal representation. These new facilities enable the expert to: i) look at the pattern highlighted on the molecule

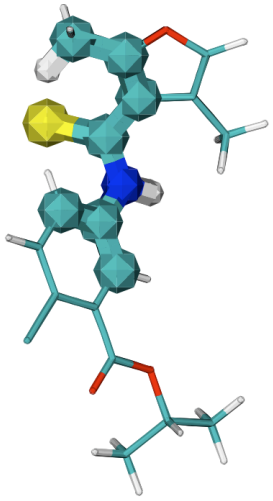

Fig. 2. HIV Pattern (wider atoms and bonds) discovered by ILP

structure; ii) interact with the visualisation tool and specify constraints not satisfied by the pattern presented; and rerun the ILP system with the specified constraints added to the data set. These steps are the centre of the main loop of the interaction where the expert guides the process of patterns discovery. Additionally the tool also allows the expert user to specify a list of chemical structures (rings and functional groups) that are used as the macro operators as described above. The use of chemical structures may be very useful to achieve more compact and comprehensible models than the ones described with atoms and bonds.

4 Conclusions

This paper reports on extensions of an existing tool to help experts in drug design tasks. The extensions include the possibility of direct guidance by the expert over the data analysis process. According to experts the extensions introduced are very useful for the drug design activity.

Acknowledgements

This work has been partially supported by the project ILP-Web-Service (PTDC-/EIA/70841/2006) and by Fundação para a Ciência e Tecnologia. Nuno A. Fonseca is funded by FCT grant SFRH/BPD/26737/2006. Max Pereira is funded by FCT grant SFRH/BPD/37087/2007.

References

1. S. Muggleton and L. De Raedt Inductive Logic Programming: Theory and Methods. 19/20:629-679, 1994.
2. J. M Collins. The DTP AIDS antiviral screen program, 1999
3. William Humphrey, Andrew Dalke, and Klaus Schulten. VMD - Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33-38, 1996
4. G. M. Maggiora, V. Shanmugasundaram, M. J. Lajiness, T. N. Doman and M. W. Schultz. A practical strategy for directed compound acquisition, pages 315-332. Wiley-VCH, 2004.
5. Stefan Kramer, Luc De Raedt, and Christoph Helma. Molecular feature mining in HIV data. In *KDD'01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 136-143, NY, USA, 2001.
6. A. Karwath and Luc De Raedt. Predictive Graph Mining In *Discovery Science, 7th International Conference, (DS 2004), Italy*, volume 3245 of *LNCS*, pages 1-15. Springer, 2004.
7. C. Borgelt and M. R. Berthold. Mining Molecular Fragments: Finding Relevant Substructures of Molecules. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), Japan*, pages 51-58, 2002.