Finding frequent substructures in chemical compounds

Luc Dehaspe
Dept. of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200A
B–3001 Heverlee, Belgium
Luc.Dehaspe@cs.kuleuven.ac.be

Hannu Toivonen
Rolf Nevanlinna Institute &
Dept. of Computer Science
P.O. Box 4, FIN–00014
University of Helsinki, Finland
Hannu.Toivonen@rni.helsinki.fi

Ross Donald King
Dept. of Computer Science
The Univ. of Wales, Aberystwyth
Penglais, Aberystwyth, Ceredigion
SY23 3DB, Wales, United Kingdom
rdk@aber.ac.uk

Abstract

The discovery of the relationships between chemical structure and biological function is central to biological science and medicine. In this paper we apply data mining to the problem of predicting chemical carcinogenicity. This toxicology application was launched at IJCAI'97 as a research challenge for artificial intelligence. Our approach to the problem is descriptive rather than based on classification; the goal being to find common substructures and properties in chemical compounds, and in this way to contribute to scientific insight. This approach contrasts with previous machine learning research on this problem, which has mainly concentrated on predicting the toxicity of unknown chemicals. Our contribution to the field of data mining is the ability to discover useful frequent patterns that are beyond the complexity of association rules or their known variants. This is vital to the problem, which requires the discovery of patterns that are out of the reach of simple transformations to frequent itemsets. We present a knowledge discovery method for structured data, where patterns reflect the one-to-many and many-to-many relationships of several tables. Background knowledge, represented in a uniform manner in some of the tables, has an essential role here, unlike in most data mining settings for the discovery of frequent patterns.

Introduction

The toxicology evaluation problem. In this paper we apply a data mining method to the problem of predicting whether chemical compounds are carcinogenic or not. This problem is of clear scientific and medical interest. Cancer is the second most common cause of death in western countries. Currently around one third of the population will get cancer sometime in their lifetime: and one quarter will die of cancer. A large percentage of these cancers are linked to environmental factors such as exposure to carcinogenic chemicals (estimated as high as 80%). Very few compounds have been fully tested for carcinogenesis as the process is very expensive and time consuming. Better computer-based methods are therefore valuable.

Copyright ©1998, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

The National Toxicity Program of the U.S. National Institute for Environmental Health Sciences conducts standardized bioassays of chemicals on rodents, in order to estimate their carcinogenetic effects on humans. Within the Predictive Toxicology Evaluation (PTE) project (Bristol, Wachsman, & Greenwell 1996) they have published a collection of chemicals already analyzed — roughly half of which have turned out to be carcinogenic — and a collection of chemicals whose tests are undergoing. Assays of the published but untested chemicals will be completed by PTE during this year. These cases offer a possibility for true blind trials in carcinogenicity prediction research. The prediction of rodent chemical carcinogenesis was launched at IJCAI '97 as a research challenge for artificial intelligence (Srinisavan *et al.* 1997). The problem is suitable for data mining as there exists a large database of chemicals available for analysis, and new knowledge needs to be discovered concerning the molecular mechanisms of carcinogenesis.

Rather than competing with expert chemists in classifying chemicals to carcinogenic or otherwise, our goal was to discover frequent patterns that would aid chemists – and data miners seeking predictive theories – to identify useful substructures for carcinogenicity research, and so contribute to the scientific insight. This can be contrasted with previous machine learning research in this application, which has mainly concentrated on predicting the toxicity of unknown chemicals (Srinisavan *et al.* 1997; Kramer, Pfahringer, & Helma 1997). We believe that a repository of frequent substructures and their frequencies would be valuable for chemical (machine learning) research. For example, once we know *all* frequent substructures, we can make stronger claims about the (non-)existence of high quality single rules than can usually be done with classifying approaches based on heuristic search.

Contribution to data mining. The task of discovering recurrent patterns has been studied in a variety of data mining settings. In its simplest form, known from association rule mining (Agrawal, Imielinski, & Swami 1993), the task is is to find all frequent itemsets, i.e., to list all combinations of items that are found in a suffi-

cient number of examples. A prototypical application example is in market basket analysis: find out which products tend to be sold together.

Our contribution to the field of data mining is in considering the discovery of useful frequent patterns that are far more complex than association rules or their known variants. We discover queries in first-order logic that succeed with respect to a sufficient number of examples. Such patterns are out of the reach of simple transformations to frequent itemsets. We present an attempt for knowledge discovery in structured data, where patterns reflect the one-to-many and many-to-many relationships of several tables. Background knowledge, represented in a uniform manner, has an essential role here, unlike in most data mining settings for the discovery of frequent patterns.

Datalog concepts. We use DATALOG (see, e.g., (Ullman 1988)) to represent both data and patterns. In DATALOG, a *term* is defined as a constant symbol, written in lowercase, or a variable, written with initial uppercase. A *logical atom* is an m-ary predicate symbol followed by a bracketed m-tuple of terms. A *definite clause* is a universally quantified formula of the form $B \leftarrow A_1, \ldots, A_n$ ($n \geq 0$), where B and the A_i are logical atoms. This formula can be read as "B if A_1 and ... and A_n". If $n = 0$, a definite clause is also called a *fact*. A (deductive) DATALOG *database* is a set of definite clauses. A formula $\leftarrow A_1, \ldots, A_n$ without a conclusion part is called a *denial*. Such a formula can also be viewed as a (PROLOG) *query ?-* $A_1, \ldots, A_n$: (the resolution based derivation of) the answer to a given query with variables $(X_1, \ldots, X_m)$ binds these variables to terms $(a_1, \ldots, a_m)$, such that the query succeeds if each X_i is replaced by a_i. This binding is denoted by $(X_1/a_1, \ldots, X_m/a_m)$. Due to the nondeterministic nature of the computation of answers, a single query Q may result in many bindings. We will refer by $answerset(Q, D)$ to the set of all bindings obtained by submitting query Q to a DATALOG database D.

Data and background knowledge. The DATALOG database for the carcinogenesis problem was taken from *http://www.comlab.ox.ac.uk/oucl/groups/machlearn/PTE/*. The set we have used contains 337 compounds, 182 (54%) of which have been classified as carcinogenic and the remaining 155 (46%) otherwise.

Each compound is basically described as a set of atoms and their bond connectivities, as proposed in (King *et al.* 1996). The atoms of a compound are represented as DATALOG facts such as *atom(d1,d1_25,h,1,0.327)* stating that compound *d1* contains atom *d1_25* of element *h* and type *1* with partial charge *0.327*. For convenience, we have defined additional view predicates *atomel*, *atomty*, and *atomch*; e.g., *atomel(d1,d1_25,h)*. Bonds between atoms are defined with facts such as *bond(d1,d1_24,d1_25,1)*, meaning that in compound *d1* there is a bond between atoms *d1_24* and *d1_25*, and the bond is of type *1*. There are roughly 18500 of these atom/bond facts to represent the basic structure of the compounds.

In addition, background knowledge contains around 7000 facts and some short DATALOG programs to define mutagenic compounds, genotoxicity properties of compounds, generic structural groups such as alcohols, connections between such chemical groups, tests to verify whether an atom is part of a chemical group, and a family of structural alerts called *Ashby* alerts (Ashby & Tennant 1991).

Representation of substructures. The target patterns or substructures are expressed as DATALOG queries. For instance, ?- *atomel(C,A,c), methyl(C,S), occurs_in(A,S)* is a pattern representing a carbon atom A that occurs in a methyl structure S within compound C.

Related work. Related problems in structure discovery in molecular biology have been considered, e.g., in (Wang *et al.* 1997; Kramer, Pfahringer, & Helma 1997; King *et al.* 1996; King & Srinivasan 1996). Substructure discovery and the utilization of background knowledge have been discussed in (Djoko, Cook, & Holder 1995). Discovery of logical patterns, similar to DATALOG queries, has been considered in (De Raedt & Dehaspe 1997) and in the context of metaqueries (Shen *et al.* 1996); they also emphasize the use of language bias for the specification of the search space. Closely related data mining problems have recently arisen also in schema discovery in semi-structured data (Wang & Liu 1997).

The substructure discovery problem we look at has a complexity somewhere between frequent itemset discovery and a full-scale inductive logic programming (ILP; see, e.g., (Lavrač & Džeroski 1994)) approach. In data mining, related problems in the area of discovering frequent patterns include association rules (Agrawal *et al.* 1996), episodes in sequences (Mannila, Toivonen, & Verkamo 1997), and sequential patterns (Agrawal & Srikant 1995), a family of problems discussed in more general in (Mannila & Toivonen 1997).

Within ILP, a closely related problem is the discovery of queries in first order logic that succeed with respect to a sufficient number of examples (Dehaspe & De Raedt 1997). In (Dehaspe & Toivonen 1998) we discuss the relationship of ILP to frequent pattern discovery, and relate data mining problems to ILP. The logical setting for substructure discovery is based on the learning from interpretations paradigm introduced in (De Raedt & Džeroski 1994).

Frequent substructure discovery

Discovery task. Intuitively, the problem we consider is the following: given the above data on chemical compounds and their structures and properties, find recur-

rent compound substructures and properties. Since the properties are also a result of the structure of a compound, for the rest of the paper we just talk collectively about (sub)structure discovery.

This problem is an instance of the generic problem of finding all potentially interesting sentences (Mannila & Toivonen 1997). Given a database $\mathbf{r}$, a class $\mathcal{L}$ of sentences (patterns), and a selection predicate q which is used for evaluating whether a sentence $Q \in \mathcal{L}$ defines a potentially interesting pattern in $\mathbf{r}$. The task is to find the theory of $\mathbf{r}$ with respect to $\mathcal{L}$ and q, i.e., the set $\mathcal{T}h(\mathcal{L}, \mathbf{r}, q) = \{Q \in \mathcal{L} \mid q(\mathbf{r}, Q) \text{ is true}\}$. In (Dehaspe & Toivonen 1998), this framework has been used to formulate the task of frequent query discovery in DATALOG. We now define frequent substructure discovery as a special case of frequent query discovery.

Definition 1 (Frequent substructure discovery) *Assume*

- $\mathbf{r}$ *is a* DATALOG *database of chemical compounds, their structures and properties, as described above,*
- $\mathcal{L}$ *is a set of substructures expressed as* DATALOG *queries* ?- $\mathrm{A}_1, \ldots, \mathrm{A}_n$, *where each logical atom* A_i *concerns some structural property of the compounds, as described above,*
- $q(\mathbf{r}, Q)$ *is true if and only if the frequency of query* $Q \in \mathcal{L}$ *with respect to* $\mathbf{r}$ *is at least equal to the frequency threshold specified by the user.*

The task is to find the set $\mathcal{T}h(\mathcal{L}, \mathbf{r}, q)$ *of all frequent substructures.*

We next define what frequency exactly means in this setting.

Definition 2 (Substructure frequency) *Given* $\mathbf{r}$ *and* $Q \in \mathcal{L}$ *as above, a relation* keypred(C), *where* keypred *is a predicate name not used in* Q *or* $\mathbf{r}$, *and* C *is the key variable used in* Q *to refer to the compound name, the (absolute) frequency of query* Q *w.r.t.* $\mathbf{r}$ *is*

$$|answerset(\text{?- keypred(C)}, \mathbf{r} \cup \{\text{keypred(C)} \leftarrow Q\})|,$$

i.e., the number of bindings of the C *variable with which the query* Q *is true in* $\mathbf{r}$, *i.e., the number of compounds in which substructure* Q *occurs.*

Frequent probabilistic rules. Once frequent substructures and their frequencies are discovered, probabilistic rules can be produced, much like in the case of association rules. In terms of the DATALOG concepts introduced above, a *probabilistic rule* R is an expression of the form $A_1, \ldots, A_k \Rightarrow A_{k+1}, \ldots, A_n$, where A_i are atoms. This formula should be read as "if query ?- $A_1, \ldots, A_k$ succeeds then query ?- $A_1, \ldots, A_n$ succeeds also". The *confidence* of rule R can be computed as the ratio of the frequencies of queries ?- $A_1, \ldots, A_n$ and ?- $A_1, \ldots, A_k$. The *frequency* (or *support*) of rule R is the frequency of query ?- $A_1, \ldots, A_n$.

The patterns searched for are significantly different from frequent itemsets or association rules, and we cannot see any useful or meaningful way of transforming the problem into a search for association rules. As an example, consider the discovery of a pattern such as

> **if** (in a given compound) a carbon atom has a bond with an atom with partial charge less than -0.2
> **then** (the compound is) carcinogenic.

As a (propositional) association rule, it seems impossible to break the condition part to a conjunction of several items. If "carbon atom" is one item and "atom with partial charge less than -0.2" another item, how can we represent the fact that they have a bond? The straightforward way is to consider the condition of the rule as just one item "a carbon atom that has a bond with an atom with partial charge less than -0.2". Such an approach with association rules leads to two serious problems.

- The number of different items explodes, as each possible combination of properties that can be associated with an atom must be considered as a separate item. Since several atoms can be related, as in the example above, the number of possible items explodes further.
- The search (e.g., with APRIORI) is inefficient as the search space is flattened and the structural information is lost. The generalization hierarchy between the derived items could in principle be treated as an item hierarchy, but the proposed methods (Han & Fu 1995; Holsheimer *et al.* 1995; Srikant & Agrawal 1995) are not feasible for complex cases like this.

As a probabilistic rule, in turn, the pattern can be expressed in a natural way as *atomel(C,A1,c), atomch(C,A2,X), -0.2 > X, bond(C,A1,A2,Y)* $\Rightarrow$ *carcinogenic(C)*. In the following section we describe WARMR, a method for searching such frequent patterns.

Substructure discovery with WARMR

We now briefly describe the WARMR algorithm used in the experiment. More details can be found in (Dehaspe & De Raedt 1997; Dehaspe & Toivonen 1998). WARMR is the first general purpose ILP system to employ the efficient levelwise method known from the APRIORI algorithm (Agrawal *et al.* 1996). In (Dehaspe & Toivonen 1998) we show how WARMR can be tuned to simulate APRIORI and some other well-known algorithms for frequent pattern discovery. A stand-alone version of WARMR is freely available for academic purposes upon request.

Language bias. WARMR is, in principle, capable of discovering arbitrary frequent DATALOG queries from a given database. In practice, however, the application domain and problem setting constrain the set of meaningful and useful patterns. In WARMR, the set of allowed patterns is specified with a declarative language bias. This well-known ILP mechanism is valuable for data mining systems: the search space is made explicit, and modifying it is easy.

The language bias is specified using expressions of the form $rmode(n : (A_1, \ldots, A_n))$, where the A_i are conjuncted logical atoms. The *rmode* declarations indicate which logical atoms can be included in a query, the maximal number of times ($n > 0$) they can be included, and the modes and types of the variables in the logical atoms. A variable V in input mode, denoted with $+V$, has to occur somewhere to the left in the query. Typing of variables can be used to constrain the occurrence of input variables, such that for instance *atomty(C,S,1)* will not be added to ?- *atomel(C,A,c), methyl(C,S)*, but *atomty(C,A,1)* or *occurs_in(A,S)* will. This format for the definition of the language of pattern $\mathcal{L}$ was originally proposed for PROGOL (Muggleton 1995) and later adapted to TILDE (Blockeel & De Raedt 1998).

Levelwise search. The levelwise algorithm (Mannila & Toivonen 1997) is based on a breadth-first search in the lattice spanned by a specialization relation $\preceq$ between patterns, where $p1 \preceq p2$ denotes pattern "$p1$ is more general than pattern $p2$", or "$p2$ is more specific than pattern $p1$". The specialisation relation used in WARMR is θ-subsumption, a strictly stronger variant of the subset relation: $p1$ θ-subsumes a $p2$ if and only if there exists a (possibly empty) binding of the variables of $p1$, such that every logical atom of the resulting query occurs in $p2$.

The levelwise method looks at a level of the lattice at a time, starting from the most general patterns. The method iterates between candidate generation and candidate evaluation phases: in *candidate generation*, the lattice structure is used for pruning non-frequent patterns from the next level; in the *candidate evaluation* phase, frequencies of candidates are computed with respect to the database. Pruning is based on monotonicity of $\preceq$ with respect to frequency: if a pattern is not frequent then none of its specialisations are frequent. So while generating candidates for the next level, all the patterns that are specialisations of infrequent patterns can be pruned.

The levelwise approach has two crucial useful properties (Mannila & Toivonen 1997). First, the database is scanned at most $k+1$ times, where k is the maximum level (size) of a frequent pattern. All candidates of a level are tested in single database pass. This is an important factor when mining large databases. Second, the time complexity is linear in the size of the result times the number of examples, assuming matching patterns against the data is fast.

Candidate generation in WARMR. To generate candidates, WARMR employs a classical specialisation operator under θ-subsumption (Lavrač & Džeroski 1994). A specialisation operator ρ maps queries $\in \mathcal{L}$ onto sets of queries $\in 2^{\mathcal{L}}$, such that for any $Query1$ and $\forall\, Query2 \in \rho(Query1)$, $Query1$ θ-subsumes $Query2$. The operator used in WARMR essentially adds conjunctions to the query as allowed by the language bias specifications.

The language bias, in particular mode declarations and the fact that conjunctions of several logical atoms can be added in a single refinement step, complicates pruning significantly. We can no longer require that all subsets of a candidate are frequent, cf. for instance APRIORI, as some of the subsets might simply not be in $\mathcal{L}$. Instead, WARMR requires candidates not to θ-subsume any infrequent query.

Candidate evaluation and memory management in WARMR. In the candidate evaluation phase the frequencies of a set of queries are computed in a single database pass. Therefore, a straightforward execution of Definition 2 (frequency of a single query) is not practical, as it would require one pass per candidate. The WARMR algorithm rather considers one compound C at a time and for each candidate Q_i runs the query ?- *keypred(C)* in database $\mathbf{r}^C \cup \mathbf{r}^{BG} \cup \{keypred(C) \leftarrow Q_i\}$, where $\mathbf{r}^C \subseteq \mathbf{r}$ only contains clauses about compound C and $\mathbf{r}^{BG} \subseteq \mathbf{r}$ is a fixed portion of background knowledge which contains clauses relevant for all compounds. If query Q_i succeeds, an associated counter q_i is incremented.

Here we take advantage of the fact that database $\mathbf{r}$ of n compounds $C_1, \ldots, C_n$, and background knowledge BG can be partitioned into $n+1$ databases $\mathbf{r}^{C_1} \cup \ldots \cup \mathbf{r}^{C_n} \cup \mathbf{r}^{BG}$. For very large databases of compounds it is essential that $\mathbf{r}^{C_i}$ is typically very small compared to $\mathbf{r}$, and can be loaded in main memory even if $\mathbf{r}$ cannot. This has the crucial advantage that evaluation of candidates Q_i can be done (relatively) efficiently with respect to a single compound at a time.

Experiments

We randomly split the set of 337 compounds into 2/3 for the discovery of frequent substructures, and 1/3 for the validation of derived probabilistic rules about carcinogenicity.

Frequent substructures. In order to investigate the usefulness of different types of information in the biochemical database, WARMR's language bias was varied to produce three sets of frequent patterns.

- *Experiment 1:* only atom element, atom type, and bond information. At level 6, WARMR generates substructure ?- *atomel(C,A1,c), bond(C,A1,A2,BT), atomel(C,A2,c), atomty(C,A2,10), atomel(C,A3,h), bond(C,A2,A3,BT)*, i.e., "a carbon atom bound to a carbon atom of type 10 bound to a hydrogen atom, where the two bonds are of the same bond type". In the training set, this highly frequent substructure is encountered in 128 compounds (57%).
- *Experiment 2:* everything except the atom/bond information. An example of a substructure discovered at level 4 is ?- *six_ring(C,S1), alcohol(C,S2),*

	E1 (F = 10%)		**E2** (F = 4%)		**E3** (F = 10%)	
L	NOC	NOFS	NOC	NOFS	NOC	NOFS
1	6	6	58	41	85	49
2	123	34	1093	413	1466	501
3	214	137	3381	2631	3219	2184
4	813	672	15411	13963	7190	6219
5	4133	3725	-	-	15577	14435
6	25434	23961	-	-	-	-
Total	29993	28535	19934	17048	27537	23388

Table 1: Results of three runs with WARMR on carcinogenicity analysis. Legend: E = experiment, F = frequency threshold, L = level, NOC = number of candidates, NOFS = number of frequent substructures.

ashby_alert(C,di10,S3), connected(S1,S3) (frequency: 11 compounds (5 %)), i.e., "an alcohol and a six ring connected to a structure with Ashby alert di10".

- *Experiment 3:* the full database, except the Ashby alerts. At level 5, WARMR produces substructure ?- *six_ring(C,S), atomel(C,A1,h), atomel(C,A2,c), bond(C,A1,A2,X), occurs_in(A2,S)* (frequency: 157 compounds (70%)), i.e., "a hydrogen atom bound to a carbon atom in a six ring".

The number of candidates and frequent patterns produced during these experiments is tabulated in Table 1. Notice that, overall, there are few infrequent candidates, and the number of candidates steadily increases. As a consequence, the exploration of deeper levels is problematic. The empty cells in the table indicate at which level the experiments were interrupted.

Probabilistic rules. As described above, our repository of frequent substructures can be exploited directly, i.e. without going back to the database, to produce probabilistic rules about carcinogenicity. For instance, we can combine ?- *cytogen_ca(C,n), sulfide(C,S)* (frequency: 7%) and ?- *non_carcinogenic(C), cytogen_ca(C,n), sulfide(C,S)* (frequency: 6%) to generate the probabilistic rule

cytogen_ca(C,n), sulfide(C,S) $\Rightarrow$ *non_carcinogenic(C)* (frequency: 6%; confidence: 86%).

To rank these rules we have applied a binomial test that verifies how unusual the confidence of rule *substructure(C)* $\Rightarrow$ *(non_)carcinogenic(C)* is, i.e. how much it deviates from the confidence of *true* $\Rightarrow$ *(non_)carcinogenic(C)*. All rules with significance below $3 * \sigma$ were discarded, with σ an estimation of the standard deviation. For instance, the significance level of the above rule is $3.16 * \sigma$. The 215 rules that passed this test were further annotated with their significance level on the 1/3 validation set, and finally combined with human domain expertise. The list below summarizes the main findings.

- In experiment 1, only using atom-bond information, no substructure described with less than 7 logical atoms is found to be related to carcinogenicity. This places a lower limit on the complexity of rules that are based exclusively on chemical structure.
- For experiments 2 and 3, validation on an independent test set showed that the rules identified as interesting in the training set were clearly useful in prediction. The estimated accuracies of the rules from the training data were optimistically biased, as expected.
- The rules found in experiments 2 and 3 are dominated by biological tests for carcinogenicity. It is very interesting that these tests appear broadly independent of each other, so that if a chemical is identified as a possible carcinogen by several of these tests, it is possible to predict with high probability that it is a carcinogen - unfortunately, such compounds are rare.
- Inspection of the rules from experiment 2 revealed that the Ashby alerts were not used by any rules. We believe this reflects the difficulty humans and machine have in discovering general chemical substructures associated with carcinogenicity - however, it is possible that the intuitive alerts used by Ashby were incorrectly interpreted and encoded in PROLOG by (King & Srinivasan 1996).
- Inspection of the rules from experiment 3 revealed no interesting substantial chemical substructures (atoms connected by bonds) in the rules found.
- Two particularly interesting rules that combine biological tests with chemical attributes were found. It is difficult to compare these with directly existing knowledge, because most work on identifying structural alerts has been based on alerts for carcinogenicity, while both rules identify alerts for non-carcinogenicity. It is reasonable to search for non-carcinogenicity alerts as there can be specific chemical mechanisms for this, e.g. cytochromes specifically neutralise harmful chemicals. The rule ?- *cytogen_ca(C,n), ring(sulfide,A,B)* for identifying non-carcinogenic compounds is interesting. The combination of conditions in the rule seems to be crucial: the cytogen and sulfide tests in isolation seem to do worse. Within rule ?- *atomch(C,A,X), X*≤ -0.215, *salmonella(C,n)* the addition of the chemical test makes the biological test more accurate at the expense of less coverage. As the rule refers to charge this rule may be connected to transport across cell membranes.

Discussion. It is interesting and significant that no atom-bond substructures described with less than 7 conditions were found to be related to carcinogenicity. This result is not inconsistent with the results obtained by (King & Srinivasan 1996) and (Srinivasan *et al.* 1997) using PROGOL because most of the substructures there involve partial charges, and the ones that don't do not meet the coverage requirements in experiment 1.

The hypothesis space which PROGOL searched to form its theory (a single complex disjunctive "alert") is larger than the hypothesis space of queries searched by WARMR. Comparing the PROGOL theory on this split it is interesting to see that the significance score is not very good on train and test set, whereas the accuracy is good on the test set, and the significance is good on the overall set.

Although the lack of significant atom-bond substructures found in experiment 1 is disappointing, it is perhaps not too surprising. The causation of chemical carcinogenesis is highly complex with many separate mechanisms involved. Therefore predicting carcinogenicity differs from standard drug design problems, where there is normally only a single well defined mechanisms. We consider that it is probable that the current database is not yet large enough to provide the necessary statistical evidence required to easily identify chemical mechanisms. Biological tests avoid this problem because they detect multiple molecular mechanisms; e.g., the Ames test for mutagenesis detects many different ways chemicals can interact with DNA and cause mutations; biological tests also detect whether the compound can cross cell membranes and not be destroyed before reaching DNA. Biological tests vary in expense, speed, and accuracy. At the extreme cheap and fast and relatively inaccurate end is the Ames test for mutagenicity, this is fast and uses bacteria (so there are no ethical issues). At the other end are long expensive trials which involve the dissection of thousands of rodents.

The ultimate goal of our work in predictive toxicology is to produce a program that can predict carcinogenicity *in humans* from just input chemical structure. Such a system would allow chemicals to be quickly and cheaply tested without harm to any animals. This goal is still far distant. Our results suggest that an intermediate goal for data mining in this predictive toxicology problem is to identify the combination of biological tests and chemical substructures that provides the most cost-effective tests for testing chemical carcinogenesis.

Conclusions

We presented a data mining problem in a biochemical database. The goal is to discover frequent substructures of chemical compounds in relation to their possible carcinogenicity. Rather than trying to predict the toxicity of unknown compounds, our purpose is to assist chemical experts in discovering chemical mechanisms of toxicology.

One result of our experiment is a repository of frequent substructures in a general DATALOG format. We believe this repository constitutes a new description of the data that is useful for chemists and data miners looking for predictive theories. We have also identified substructures, both known and new, that could be related to carcinogenicity. On the other hand, we have found that, within this biochemical database, short, accurate and highly significant rules apparently do not exist.

The frequent substructures that we search for are described as DATALOG queries. We gave examples of such patterns found in the carcinogenicity database, and we argued that association rules are not a suitable representation for patterns needed in this application. Relational patterns are needed instead to describe useful aspect of substructures and their properties.

Although we have discussed one application only in this paper, the problems attacked are general. While association rules are a useful formalism, there are a number of problems where much more expressive frequent patterns can be useful. In the domain of market basket analysis, for instance, a number of properties can be associated with products, such as the department, the price, etc. With association rules it would be difficult to express, e.g., the pattern that if something in promotion is purchased then something else is purchased from the same department. Or, consider telecommunication alarm analysis and the discovery of episodes (Mannila, Toivonen, & Verkamo 1997) and a frequent pattern of the form

> **if** alarm of class c from device X and Y is a device connected to X and Y is a base station **then** alarm of class d from device Y.

Such patterns cannot be discovered with a straightforward application of propositional association or episode rule discovery methods. Specialized solutions can certainly be found for specific cases, but with a loss of generality and flexibility.

In this paper we outlined the WARMR method for discovering frequent DATALOG queries. We would like to point out two useful properties of the method. First, by changing the language bias the user can easily search for different patterns without modifying the algorithm. Second, WARMR can utilize background knowledge specified as tables or DATALOG programs. In both these respects WARMR differs favourably from most data mining methods for mining frequent patterns.

Acknowledgements. Luc Dehaspe is supported by ESPRIT Long Term Research Project No 20237, ILP^2. Hannu Toivonen is supported by the Academy of Finland. R.D. King was partly supported by grant BIF08765 from the BBSRC and grant GR/L262849 from the EPSRC. We are grateful to Luc De Raedt and Heikki Mannila for discussions and comments, and to Wim Van Laer and Hendrik Blockeel for their share in the implementation of WARMR.

References

Agrawal, R., and Srikant, R. 1995. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering (ICDE'95)*, 3 – 14.

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; and Verkamo, A. I. 1996. Fast discovery of association rules. In Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds., *Advances in Knowledge*

Discovery and Data Mining. Menlo Park, CA: AAAI Press. 307 – 328.

Agrawal, R.; Imielinski, T.; and Swami, A. 1993. Mining association rules between sets of items in large databases. In Buneman, P., and Jajodia, S., eds., *Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD'93)*, 207 – 216. Washington, D.C., USA: ACM.

Ashby, J., and Tennant, R. W. 1991. Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. *Mutation Research* 257:229–306.

Blockeel, H., and De Raedt, L. 1998. Top-down induction of first order logical decision trees. *Artificial Intelligence*. To appear.

Bristol, D.; Wachsman, J.; and Greenwell, A. 1996. The NIEHS predictive-toxicology evaluation project. *Environmental Health Perspectives* Supplement 3:1001–1010.

De Raedt, L., and Dehaspe, L. 1997. Clausal discovery. *Machine Learning* 26:99–146.

De Raedt, L., and Džeroski, S. 1994. First order jk-clausal theories are PAC-learnable. *Artificial Intelligence* 70:375–392.

Dehaspe, L., and De Raedt, L. 1997. Mining association rules in multiple relations. In *Proceedings of the 7th International Workshop on Inductive Logic Programming*, volume 1297 of *Lecture Notes in Artificial Intelligence*, 125–132. Springer-Verlag.

Dehaspe, L., and Toivonen, H. 1998. Frequent query discovery: a unifying ILP approach to association rule mining. Technical Report CW-258, K.U.Leuven. *http://www.cs.kuleuven.ac.be/publicaties/rapporten/CW1998.html*.

Djoko, S.; Cook, D. J.; and Holder, L. B. 1995. Analyzing the benefits of domain knowledge in substructure discovery. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD'95)*, 75 – 80.

Han, J., and Fu, Y. 1995. Discovery of multiple-level association rules from large databases. In *Proceedings of the 21st International Conference on Very Large Data Bases (VLDB'95)*, 420 – 431.

Holsheimer, M.; Kersten, M.; Mannila, H.; and Toivonen, H. 1995. A perspective on databases and data mining. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD'95)*, 150 – 155. Montreal, Canada: AAAI Press.

King, R., and Srinivasan, A. 1996. Prediction of rodent carcinogenicity bioassays from molecular structure using inductive logic programming. *Environmental Health Perspectives* 104(5):1031–1040.

King, R.; Muggleton, S.; Srinivasan, A.; and Sternberg, M. 1996. Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. *Proceedings of the National Academy of Sciences* 93:438–442.

Kramer, S.; Pfahringer, B.; and Helma, C. 1997. Mining for causes of cancer: machine learning experiments at various levels of detail. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD'97)*, 223 – 226.

Lavrač, N., and Džeroski, S. 1994. *Inductive Logic Programming: Techniques and Applications*. Ellis Horwood.

Mannila, H., and Toivonen, H. 1997. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery* 1(3):241 – 258.

Mannila, H.; Toivonen, H.; and Verkamo, A. I. 1997. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery* 1(3):259 – 289.

Muggleton, S. 1995. Inverse entailment and Progol. *New Generation Computing* 13.

Shen, W.; Ong, K.; Mitbander, B.; and Zaniolo, C. 1996. Metaqueries for data mining. In Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds., *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press. 375–398.

Srikant, R., and Agrawal, R. 1995. Mining generalized association rules. In Dayal, U.; Gray, P. M. D.; and Nishio, S., eds., *Proceedings of the 21st International Conference on Very Large Data Bases (VLDB'95)*, 407 – 419. Zürich, Switzerland: Morgan Kaufmann.

Srinisavan, A.; King, R. D.; Muggleton, S. H.; and Sternberg, M. J. E. 1997. The predictive toxicology evaluation challenge. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97)*. Morgan Kaufmann.

Srinivasan, A.; King, R.; Muggleton, S.; and Sternberg, M. 1997. Carcinogenesis predictions using ILP. In *Proceedings of the 7th International Workshop on Inductive Logic Programming*, Lecture Notes in Artificial Intelligence, 273–287. Springer-Verlag.

Ullman, J. D. 1988. *Principles of Database and Knowledge-Base Systems*, volume I. Rockville, MD: Computer Science Press.

Wang, K., and Liu, H. 1997. Schema discovery for semistructured data. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD'97)*, 271 – 274.

Wang, X.; Wang, J. T. L.; Shasha, D.; Shapiro, B.; Dikshitulu, S.; Rigoutsos, I.; and Zhang, K. 1997. Automated discovery of active motifs in three dimensional molecules. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD'97)*, 89 – 95.