Journal of Computer-Aided Molecular Design, **15:** 173–181, 2001.
KLUWER/ESCOM

Warmr: a data mining tool for chemical data

Ross D. King[a], Ashwin Srinivasan[b] & Luc Dehaspe[c]
[a]*Department of Computer Science, University of Wales, Aberystwyth Penglais, Aberystwyth, Ceredigion, SY23 3DB Wales, UK;* [b]*Computing Laboratory, University of Oxford, Oxford OX1 3QD, UK;* [c]*PharmaDM, Celestijnenlaan 200A, B-3001, Belgium*

Received 28 October 1999; Accepted 9 October 2000

Key words: carcinogenesis, chemical structure, inductive logic programming, machine learning, predictive toxicology

Summary

Data mining techniques are becoming increasingly important in chemistry as databases become too large to examine manually. Data mining methods from the field of Inductive Logic Programming (ILP) have potential advantages for structural chemical data. In this paper we present Warmr, the first ILP data mining algorithm to be applied to chemoinformatic data. We illustrate the value of Warmr by applying it to a well studied database of chemical compounds tested for carcinogenicity in rodents. Data mining was used to find all frequent substructures in the database, and knowledge of these frequent substructures is shown to add value to the database. One use of the frequent substructures was to convert them into probabilistic prediction rules relating compound description to carcinogenesis. These rules were found to be accurate on test data, and to give some insight into the relationship between structure and activity in carcinogenesis. The substructures were also used to prove that there existed no accurate rule, based purely on atom-bond substructure with less than seven conditions, that could predict carcinogenicity. This results put a lower bound on the complexity of the relationship between chemical structure and carcinogenicity. Only by using a data mining algorithm, and by doing a complete search, is it possible to prove such a result. Finally the frequent substructures were shown to add value by increasing the accuracy of statistical and machine learning programs that were trained to predict chemical carcinogenicity. We conclude that Warmr, and ILP data mining methods generally, are an important new tool for analysing chemical databases.

Introduction

Large chemoinformatic databases are now commonplace. To fully exploit these databases new computer based data analysis methods are required. One class of algorithms that is well suited to this task are data mining algorithms [1, 2]. In computer science the term 'data mining' refers strictly to a class of algorithms from the field of Knowledge Discovery in databases (KDD), which aim to find interesting patterns in databases. Data mining algorithms differ from those from statistics, neural networks, and traditional symbolic machine learning, both in their emphasis on efficiency (allowing large databases to be dealt with), and their emphasis on extracting comprehensible knowledge.

The prototypical data mining task is to find all frequently occurring patterns of a particular type. In its simplest form, known as association rule mining [3], the task is to find all frequent itemsets, i.e., to list all combinations of items that are found together in a sufficient number of examples. A typical application of association rules is market basket analysis, where you identify all products which tend to be sold together – this information can then be used to influence product placement, etc. Directly translating association rule mining into a chemoinformatic context, with molecules as shopping baskets: the task would be to find all elements that occur frequently together in molecules. As this translation makes clear, the standard data mining task of association rule mining is not directly transferable to chemical databases.

What is important in chemical databases is not the frequency of co-occurrence of individual atoms, but the frequency of occurrence of particular molecular sub-structures. There are two approaches to incorporating molecular sub-structures. The standard one is to use attributes to represent structure. Attributes are descriptors which describe a property of a whole object. For example, typical attributes in chemoinformatics are: the hydrophobicity of a compound, the presence of a particular molecular substructure, the charge at a particular co-ordinate position, etc. It is a characteristic of the use of attributes that all the information about a particular example can be put into a single row of a table. The use of attributes is standard in statistics, neural networks, and machine learning [4]. It is also standard in chemoinformatics: traditional QSAR [5, 6], CASE/MULTICASE [7], CoMFA [8], Recursive Partitioning [9, 10], etc., are all based on attributes.

The alternative approach, and the one which we favour, is to use a relational language to describe chemicals. This approach is known as Inductive Logic Programming (ILP) [11, 12]. ILP has clear theoretical advantages for chemoinformatics. Chemical structure are naturally relational and they can only be approximated using attributes. (One way of viewing programs such as SCAM [10] which enumerate all possible small structural attributes is that they are approximating an ILP approach – similar to the machine learning program LINUS [12]). ILP has shown its value in many conventional structure-function problems where it has found solutions not accessible to standard statistical, neural network, or genetic algorithms [13–19]. ILP based drug design methods have been successfully extended from standard QSAR problems [13–15], to toxicology [16–18], and to pharmacophore discovery [19]. With ILP:

- There is no need for all the information about a particular example to be forced into a single row of a table – which may result in loss of information. The information may be preserved by spreading it several tables.
- Human comprehensible results are more easily produced, because the use of logical relations provides a richer language that is closer to natural language.
- There is no need to pre-align structures to an extrinsic co-ordinate system (this advantage is not applicable to the problem tackled in this paper).

The main disadvantages of ILP have been its need for large amounts of computing resources, and the specialised expertise needed to use it. Whether ILP's advantages are sufficiently strong in practice to make ILP the generally preferred methodology for a chemoinformatic problems is still an open question.

In this paper we describe the ILP data mining algorithm Warmr [20, 21]. Warmr is a general purpose ILP data mining tool that finds frequent relational patterns in databases. It has been applied to a number of different application areas, e.g. telecommunications [21]. The efficiency of Warmr scales linearly with database size and it has been applied to datasets containing many millions of data points. This answers some of the efficiency problems of ILP. The frequent relational patterns found by Warmr can be used in predictive theories and contribute to scientific insight. In this paper we apply ILP data mining to chemoinformatic data for the first time. We use Warmr to find all frequent patterns in a database of chemicals tested for carcinogenesis in rodents.

Methods

Predicting rodent carcinogenesis

We applied the ILP data mining algorithm Warmr to a database of long term carcinogencity tests of compounds in rodents. This database was formed by the US National Toxicology Program (NTP) of the National Institute of Environmental Sciences (NIEHS) [22]. The database is a good test bed to illustrate the applicability of ILP data mining to chemoinformatics as it is reasonably large (25 500 facts), and has been previously intensively studied [23–26]. Much of the work on the database has been prompted by the Predictive Toxicology Evaluation (PTE) project [25–27]. The PTE identifies sets of chemical that are scheduled or ongoing assay. These compounds are then used as 'blind test' data for predictive toxicology programs. So far two PTE trials have taken place [25, 26] and a third is planned [27]. In this paper we use the data from PTE-2 as a test set. This allows direct comparison with a wide variety of other results. The database for the carcinogenesis problem was taken from http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/PTE/. This was the official site of the International Joint Conference on Artificial Intelligence (IJCAI) challenge on predictive toxicology which was associated with NTP program on predictive toxicology evaluation PTE-2 [28, 29]. The dataset contains 337 compounds, 182 (54%) of which have been classified as carcinogenic and the remaining 155 (46%) otherwise. Each compound is described by the results of

the carcinogenicity assay: carcinogenic or not carcinogenic and a database of background knowledge about the compound. These properties are described using a special logical language:

1. Atom-bond description. This consists of a set of atoms and their bond connectivities, as described in [16] (for convenience, we have modified the original representation slightly). Atoms are described using three relations (predicates) *atom_element*, *atom_type*, and *atom_charge* for atom element type, descriptor type, and partial charge respectively. For example: *atom_element (compound1, id1_24, hydrogen)*, describes a hydrogen atom in compound 1 with identifier *id1_24; atom_type (compound1, id1_27, 10)*, describes an atom of type 10 (aliphatic carbon) in compound 1 with identifier *id1_27; atom_charge(compound1, id1_27, 0.01)*, describes an atom with a partial charge of 0.01 in compound 1 with identifier id1_27. Bonds are described using bond relations such as *bond(compound1, id1_24, d1_25, 1)*, meaning that in compound 1 there is a single bond (type 1) between atoms with identifiers *d1_24* and *1d1_25*. The motivation for using this representation is that atoms and bonds are the fundamental building blocks of chemistry, and from them all other representations are built. The approach has previously also been successful [16–19]. This data is relational.
2. Generic structural groups. This represents generic chemical structural groups (methyl, alcohol, benzene rings, etc.). These are defined in datalog (see below) and so the definitions can also be considered part of the background knowledge. We used 29 different structural groups, which expands on the 12 definitions used in our mutagenesis study [16]. An example structural relation is: *group(compound1, ether, [id1_12, id1_13, id1_14])*. This states that compound 1 has an ether group consisting of the atoms with identifiers id1_12, id1_13, and id1_14. The motivation for using chemical groups is that after atom/bonds they are the simplest way to represent chemical structure. This data is relational.
3. Genotoxicity. These are results of short-term assays used to detect and characterize chemicals that may pose genetic risks. These assays include the Salmonella assay, in-vivo tests for the induction of micro-nuclei in rat and mouse bone marrow, etc. The results are usually positive or negative. An example genotoxicity relation is: *salmonella(compound1, positive)*. This states that compound 1 had a positive result on the *Salmonella* genotoxicity assay. In cases where more than 1 set of results are available for a given type, we used the majority result. When positive and negative results are returned in equal numbers, then no result is recorded for that test. The motivation for using these properties is that they have been shown to be important in accurate prediction of carcinogenicity [22–26]. This data is attribute based.
4. Mutagenicity. Progol rules from the earlier experiments on obtaining structural rules for mutagenesis are included. Mutagenic chemicals have often been found to be carcinogenic, and we use all the rules found with Progol. An example mutagenicity relation is: *mutagenic(compound1)*. This states that compound 1 is predicted to be mutagenic. Mutagenicity is probably the most important test for determining if a chemical is carcinogenic [22–26]. This data is attribute based.
5. Structural indicators. We have encoded some structural alerts thought to be associated with carcinogenesis based on the work of Ashby and co-workers [23, 24]. An example structural indicator relation is: *indicator(compound1, nitro, [id1_30, id1_31, id1_32, id1_33])*. This states that compound 1 has an alert of type nitro involving atoms with identifiers id1_30, id1_31, id1_32, and id1_33. This data is attribute based.

The 337 compounds are described using roughly 25 500 of these logical relations. The data mining system is designed to find frequent patterns among such logical relations.

Warmr

Warmr is a general purpose Inductive Logic Programming (ILP) data mining algorithm [20, 21]. It uses *datalog* [30] to represent both data and patterns. Datalog is a logic programming language (with no function symbols) specifically designed to implement deductive databases (databases that can incorporate rules as well as facts). The relations described above are in datalog. Warmr can discover knowledge in structured data, where patterns reflect the one-to-many and many-to-many relationships of several tables. *This is not possible with standard data mining programs.* Background knowledge is represented in a uniform manner and has an essential role in the discovery of frequent patterns, unlike in most data mining settings.

Warmr used the efficient levelwise method known from the Apriori algorithm [31]. This allows it to be used on very large databases. The Warmr levelwise search algorithm [32] is based on a breadth-first search of the pattern space (Figure 1). This space is ordered by the generality of patterns. The levelwise method searches this space one level 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 the monotonicity of specificity with respect to frequency – if a pattern is not frequent then none of its specialisations can be 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 [32]. 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 with the number of examples – assuming matching patterns against the data is fast. We have previously shown how Warmr can be tuned to simulate Apriori and some other well-known algorithms for frequent pattern discovery [20]. Warmr is, in principle, capable of discovering arbitrary frequent datalog queries from a given database. However, background knowledge is generally used to constrain the set of meaningful and useful patterns. This is specified using a special control language [33, 34] which allows the search space to be made explicit, and modified easily.

Results

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

We investigated the usefulness of using three types of information by forming three databases:

- In database 1 we used only atom element, atom type, and bond information.
- In database 2 we used all the data except the atom/bond information.
- In database 3 we used all the data except the Ashby alerts.

Table 1. The results for Database 1 (using only atom and bond information). The candidates are the patterns generated by the pattern generation language. The frequency of patterns is set at 10% – this means that if a pattern has to cover 10% of the examples to be considered frequent

Level	Candidate patterns	Frequent patterns	Time (s)
1	6	6	1
2	123	34	3
3	214	127	24
4	813	672	164
5	4133	3725	3102
6	25434	23961	101673
Total	29993	28535	104957

The number of candidate patterns and the number of frequent ones are given in Tables 1–3. Notice that, overall, there are few infrequent candidates, and the number of candidates steadily increases with size of pattern. As a consequence, the exploration of complex patterns is computationally expensive. Exploration of each database was terminated when a fixed amount of computational resources were used. A complete list of all frequent substructures can be found at anonymous ftp: ftp://ftp.dcs.aber.ac.uk/pub/users/rdk/warmr/ (Note a slightly different but equivalent syntax is used).

Table 1 gives the results for database 1. The six candidate patterns of size 1 correspond to the presence of the elements: carbon, hydrogen, oxygen, nitrogen, chlorine, and sulphur (in the order of frequency of occurrence). Note that, unsurprisingly, all these patterns are frequent. An example frequent pattern of size 6 is:
atom_element(Compound, Id1, carbon) &
atom_element(Compound, Id2, carbon) &
atom_element(Compound, Id3, hydrogen) &
atom_type(Compound, Id1, 10) &
bond(Compound, Id1, Id2, Bond_type) &
bond(Compound, Id1, Id3, Bond_type).
A translation of this pattern into English is: ‘A compound with a carbon atom (Id1) of type 10 is single bonded to another carbon atom (Id2) and to a hydrogen atom (Id3)’. Names that start with an upper-case letter are variables, and those that start with a lower-case letter are constants, and that the bonds are forced to be of the same type, and Carbon type 10 is an aliphatic

Table 2. The results for Database 2 (all descriptors except the atom and bond information). The frequency of patterns is set at 4%

Level	Candidate patterns	Frequent patterns	Time (s)
1	58	41	1
2	1093	413	47
3	3381	2631	1001
4	15411	13963	17102
Total	19934	17048	18151

Table 3. The results for Database 3 (all descriptors except the Ashby alerts). The frequency of patterns is set at 10%

Level	Candidate patterns	Frequent patterns	Time (s)
1	85	49	1
2	1466	501	89
3	3219	2184	2342
4	7190	6219	194456
5	15577	14435	96896
Total	27537	23388	118784

carbon, so the bonds are single. This substructure is clearly a common one in organic compounds. In the training set this frequent substructure is found in 57% of compounds.

Table 2 gives the results for database 2. Example candidate patterns of size 1 are *group(Compound, alcohol, Group1)* and *salmonella(Compound, positive).* The first pattern describes the presence of an alcohol group, the second to the property of being positive to the salmonella mutagenicity test. An example frequent pattern of size 5 is
group(Compound, ether, Group1) &
group(Compound, ester, Group2),
group(Compound, alcohol, Group3) &
connected(Group1, Group2) &
connected(Group1, Group3).
A translation of this substructure into English is: 'A compound with an ether group connected to an ester group and an alcohol group'. In the training set this frequent substructure is found in 4% of compounds.

Table 3 gives the results for database 3. An example frequent pattern of size 4 is:
salmonella(Compound, positive) &
cytogen_sce(Compound, positive) &
cytogen_ca(Compound, positive),
group(Compound, hexane, Group1).
A translation of this into English is: A compound with a positive *Salmonella* test, a positive cytogen_sce test, a positive cytogen_ca test, and a hexane group. Note that the definition of pattern used is a database definition- objects related together in the database – and this does not only specify chemical substructures. This pattern combines together results of biological tests and the presence of chemical groups. In the training set this frequent substructure is found in 18% of compounds.

Probabilistic rules

One useful application of the repository of frequent substructures is to generate probabilistic rules. These can be generated directly, without going back to the database. For instance, we can combine the following two frequent patterns:
(1) *cytogen_ca(Compound, negative) &*
group(Compound, sulfide, Group1)
(frequency: 7%)
and
(2) *class(Compound, non_carcinogenic) &*
cytogen_ca(Compound, negative) &
group(Compound, sulfide, Group1)
(frequency: 6%)
To generate the probabilistic rule:
(3) if
cytogen_ca(Compound, negative) &
group(Compound, sulfide, Group1)
then
class(Compound, non_carcinogenic)
(frequency: 6%; confidence: 86%).

This is possible because 1 and 2 logically imply 3 (and $6/7 = 0.86$). To rank these probabilistic rules we have applied a binomial test that verifies how unusual the confidence of rule substructure is. The test compares how far the rule: if pattern then (non)carcinogenic is from the expected class frequencies. All rules with significance below 3σ were discarded (σ is the estimation of the standard deviation). For instance, the significance level of the above rule is 3.16σ. The 215 rules that passed this test were further annotated with their significance level on the $\frac{1}{3}$ validation set, and examined by eye.

In the experiment using database 1 (only using atom-bond information), *no substructure described with less than 7 logical parts was found to be related to carcinogenicity.* This result is significant because it places a lower limit on the complexity of rules that

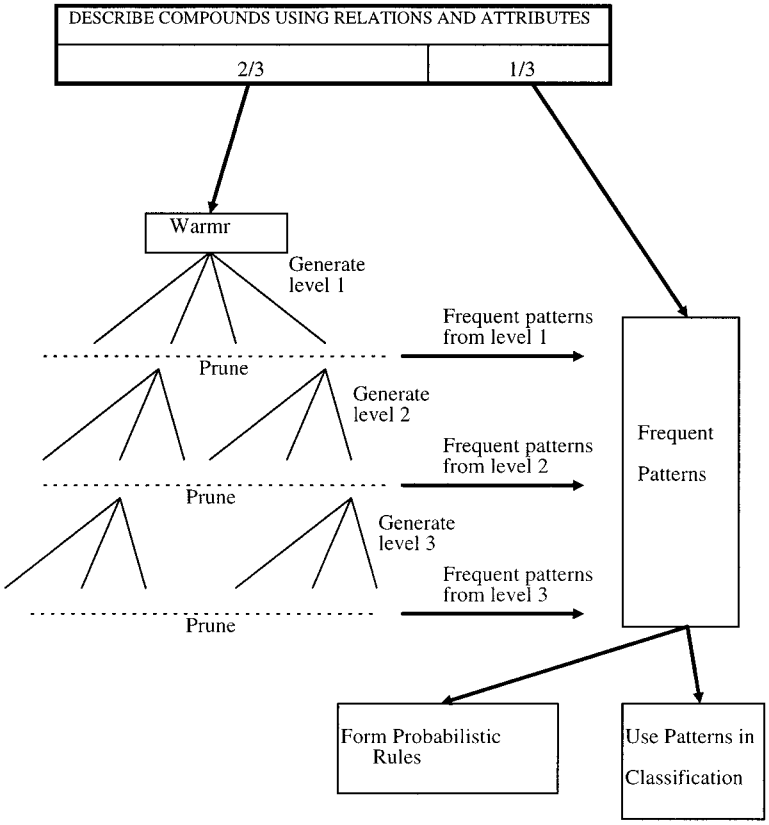

Figure 1. Data mining methodology. The input to Warmr are the descriptions of the compounds. Warmr goes through successive rounds of generating new patterns by adding to existing patterns and pruning patterns that don t occur frequently. At each level another logical condition is added to the pattern. This approach is efficient and effective because a pattern that is infrequent cannot be used to form a frequent pattern by adding conditions to it. The maximum level searched is limited by pre-set computer resources. We demonstrated the utility of the frequent patterns by forming probabilistic rules from them and by using them as attributes for classification. One third of the data was kept back to validate the probabilistic rules.

are based exclusively on atom-bond chemical structure. As we have done a complete search, we have shown that no molecular structure consisting of seven or less atoms/bonds is a good indicator of rodent carcinogenicity.

For experiments 2 and 3, validation on an independent test set showed that the frequent rules identified 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 were dominated by biological tests for carcinogenicity. The biological tests appear to be 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. This result, that the Ashby alerts provide little information, confirms that of Bahler & Bristol [35]. We believe this reflects the difficulty humans and machines 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 both [16] and [35].

Two particularly interesting rules that combine biological tests with chemical attributes were found. It is difficult to compare these with existing knowledge, as most work on identifying structural alerts has been based on alerts for carcinogenicity, while both rules identify alerts for non-carcinogenicity. However, it is reasonable to search for non-carcinogenicity alerts, as there can be specific chemical mechanisms for this,

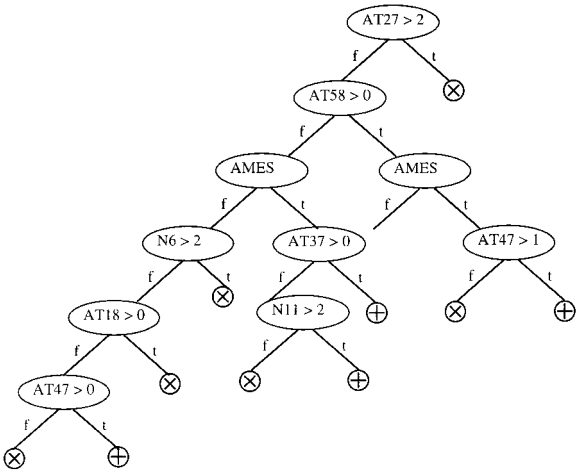

Figure 2. Decision tree formed by C4.5 without the Warmr attributes. The attribute ids are taken from (http://www.comlab.ox.ac.uk/oucl/groups/machlearn/PTE/oucl2.html). t = true. f = false. ($\otimes$ = carcinogen. ($\oplus$ = non-carcinogen. AMES – Ames test. The following attributes are based on Quanta types: AT18 - number of carbon atoms with QUANTA type 29 (indicates presence of biphenyl group); AT27 number of hydrogen atoms of Quanta type 8 (hydroxyl hydrogen); AT37 number of nitrogen atoms of Quanta type 35 (nitrogen in 6-membered ring attached to quanidinium group); AT47 number of oxygen atoms of Quanta type 49 (ester oxygen); AT58 number of sulphur atoms of Quanta type 74 (thioether) . The following attributes are based on National Toxicity Program attribute: N6 NTP descriptor 6 (RZ principal axis of inertia), N11 NTP descriptor 11 (HOMO).

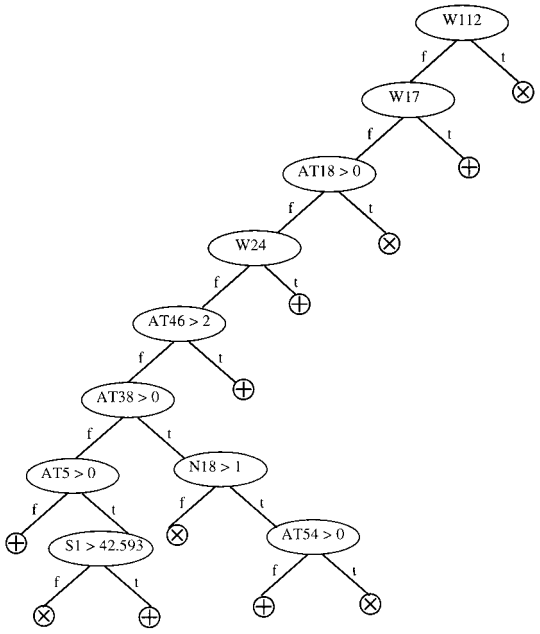

Figure 3. Decision tree formed by C4.5 with the Warmr attributes (staring with W). The attribute ids are taken from (http://www.comlab.ox.ac.uk/oucl/groups/machlearn/PTE/oucl2.html). t = true. f = false. ($\otimes$ = carcinogen. ($\oplus$= non-carcinogen. The following attributes are based on Quanta types: AT5 – number of carbon atoms of QUANTA type 14 (indicates presence of carbonyl carbon); AT18 – number of carbon atoms with QUANTA type 29 (indicates presence of biphenyl group); AT46 – umber of oxygen atoms of QUANTA type 45 (indicates presence of hydroxyl oxygen); AT54 – number of phosphorous atoms of QUANTA type 62 (indicates presence of phosphate group). S1 – NTP bulk property (the smallest principal moment of inertia). N18 – generic group count feature 18 (the number of methoxy groups in a molecule). The following attributes are Warmr generated: W17. This is true if a molecule contains at least 1 atom with partial charge at most -0.625, and tests 'n' on the genotoxicity test for Salmonella and on 'p' the Mouse Lymphoma; W24 – this is true if a molecule tests 'n' on genotoxicity test for in-vitro cytogenetics (CA) and has at least 1 sulfide group; W112 – this is true if a molecule tests 'p' on the following genotoxicity tests: Mouse Lymphoma and *Drosophila* SLRL.

e.g., cytochrome p450s specifically neutralise harmful chemicals. The rule:

> *if cytogen_ca(Compound, negative) & group (Compound, sulfide, Group1) then class(Compound, non_carcinogenic)*

is intriguing. The rule states that if a compound is negative to the cytogen_ca test and has a sulfide group then it will be non-carcinogenic. It is the combination of conditions in the rule which seems to be crucial, as the cytogen_ca test and the presence of sulfide in isolation do much worse. The cytogen_ca test is found to be particularly accurate for compounds with sulfide groups. The rule query:

> *if atom_ch(Compound, Atom, Charge) & Charge ≤ -0.215 & salmonella(Compound, negative) then class(Compound, non_carcinogenic).*

is also interesting. It states that if a compound has an atom with a partial charge ≤ -0.215 and has negative result on the Salmonella assay then the compound will be non-carcinogenic. Analysis shows that the addition of the chemical test makes the biological test more accurate at the expense of less coverage. The rule may be connected to transport across cell membranes.

Classification

To further show the utility of the repository of frequent patterns generated by Warmr, we made them available to the participants in the Predictive Toxicology Evaluation Challenge (PTE – see above) [27–29]. To do this we encoded them as new descriptive attributes of the data (indicator variables). This way of using ILP has previously been shown to be successful in preprocessing chemoinformatic data [18].

The result was that the top three most accurate methods all used the Warmr attributes in their prediction models, see: (http://www.comlab.ox.ac.uk/oucl/groups/machlearn/PTE) [29].

The results for the OUCL-2 group are illustrative of the value of using the new attributes. Figure 2 shows the decision tree learnt by the C4.5 algorithm [36] using the attributes other than those formed using Warmr. This decision tree has an estimated error rate of 33.8%. Figure 3 shows the decision tree learnt by the C4.5 algorithm including the Warmr attributes. This decision tree has an estimated error rate of 31.2%. The decision tree generated with the Warmr attributes is more accurate and somewhat simpler, as it is more linear.

Discussion and conclusion

Perhaps the most interesting result found by Warmr is that no atom-bond substructures described with less than seven conditions were found to be related to carcinogenicity. Only by using a data mining algorithm, and doing a complete search, is it possible to prove such a result. The result is consistent with the results obtained by [16] and [28] using Progol, as most of the substructures previously found involved partial charges, and the ones that didn t do not meet the coverage requirements in experiment 1. Although the lack of significant atom-bond substructures found in experiment 1 is disappointing, it is perhaps not surprising. The causation of chemical carcinogenesis is highly complex with many separate mechanisms involved.

Large databases are becoming increasingly important in chemistry. To fully exploit these databases automatic computational methods are required. Warmr is the first ILP data mining algorithm to be applied to chemoinformatic data. We illustrated the value of Warmr by finding all frequent substructures in a well studied large database. These frequent substructures were shown to add value to the database by being converted into probabilistic rules, and by being used as attributes (indicator variables) for standard prediction algorithms. These results show that ILP Data mining techniques are an important new tool for chemists.

Acknowledgements

Luc Dehaspe was supported by ESPRIT Long Term Research Project No 20237, ILP. We thank Hannu Toiven and Luc De Raedt for valuable discussions.

Availability

A stand-alone version of Warmr is freely available for academic purposes upon request (Luc.Dehaspe@cs.kuleuven.ac.be).

References

1. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy (Eds) Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA, 1996.
2. Communications of the ACM. Special issue on data mining 39, 11 (1996).
3. Agrawal, R., Imielinski, T. and Swami, A., in Buneman, P. and Jajodia, S. (Eds), Proceedings of the ACM SIGMOD Conference on Management of Data (1993) 207–216.
4. Mitchell, T.M. Machine Learning. McGraw-Hill, New York, NY, 1997.
5. Hansch, C., Malony, P.P., Fujiya, T. and Muir, R.M., Nature 194, (1962) 178.
6. Martin, Y.C. Quantitative Drug Design: A Critical Introduction, Marcel Dekker, New York, NY, 1978.
7. Klopman, G., J. Am. Chem. Soc., 106 (1984) 7315.
8. Cramer, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
9. Chen, X., Rusinko, A. and Young, S.S., J. Chem. Inf. Comput. Sci., 38 (1998) 1054.
10. Rusinko, A., Farmen, M.W., Lambert, C.G., Brown, P.L. and Young, S.S., J. Chem. Inf. Comput. Sci., 39 (1999) 1017.
11. Muggleton, S. (Ed.) Inductive Logic Programming. Academic Press, London, 1992.
12. Lavrac, N. and Dzeroski, S., Inductive Logic Programming: Techniques and Applications. Ellis Horwood, Chichester, 1994.
13. King, R.D., Muggleton, S., Lewis R.A. and Sternberg, M.J.E., Proc. Natl. Acad. Sci. USA, 89 (1992) 11322.
14. Hirst, J.D., King, R.D. and Sternberg, M.J.E., J. Comp. Aid. Mol. Des., 8 (1994) 405.
15. Hirst, J.D., King, R.D. and Sternberg, M.J.E., J. Comp. Aid. Mol. Des., 8 (1994) 421.
16. King, R.D., Muggleton, S.H., Srinivasan, A. and Sternberg, M.J.E., Proc. Natl. Acad. Sci. USA, 93 (1996) 438.
17. King, R.D. & Srinivasan, A., Env. Health Perspect., 104 (supplement 5) (1996) 1031.
18. King, R.D. and Srinivasan, A., J. Comp. Aid. Mol. Des., 11 (1998) 571.
19. Finn, P., Muggleton S., Page, D. and Srinivasan, A., Machine Learning J., 30 (1998) 241.
20. Dehaspe, L. and De Raedt, L., Lecture Notes in Artificial Intelligence, vol. 1297. Springer-Verlag, New York, NY, 1997.
21. Dehaspe, L. and Toivonen. H., Data Mining Knowledge Discovery, 3 (1999) 7.
22. Huff, J. and Hasernan, J., Env. Health Perspect., 96 (1991) 23.
23. Ashby, J. and Tennant, R.W., Mutation Res., 257 (1991) 229.
24. Tennant, R.W., Spalding, J., Stasiewicz, S. and Ashby, J., Mutagenesis, 5 (1990) 3.
25. Bahler, D.R. and Bristol, D.W., in Hunter, L., Searls, D. and Shavlik, D. (Eds), Proceedings of the First International Conference on Intelligent Systems for Molecular Biology MIT Press, Menlo Park, 1993, pp. 29–37.
26. Bristol, D.W., Wachsman, J.T. and Greenwall, A., Env. Health Perspect., 104 (supplement 5) (1996) 1001.

27. Srinivasan, A., King, R.D., Bristol, D.W., in Dzeroski, S and Flach, P.A. (Eds), Proceedings of the Ninth International Workshop on Inductive Logic programming LNAI. Springer-Verlag Berlin, 1999, pp. 291–302.
28. Srinivasan, A., King, R.D., Muggleton, S.H. and Sternberg, M.J.E., Fifteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Francisco, 1997, pp. 4–9.
29. Srinivasan, A., King, R.D., Bristol, D.W., Sixteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Francisco, 1999, pp. 270–275.
30. Ullman, J.D., Principles of Database and Knowledge-Base Systems. MD Computer Science Press, Rockville, 1988.
31. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A.I., in Fayyad, U.M, Piatetsky-Shapiro, G., Smyth and Uthurusamy, R. (Eds), Advances in Knowledge Discovery and Data mining AAAI Press, Menlo Park, CA, 1996, pp. 307–328.
32. Mannila, H. and Toivonen, H., Data Mining and Knowledge Discovery, 1 (1997) 241.
33. Muggleton, S., New Gen. Comput., 13 (1995) 245.
34. Blockeel, H. and De Raedt, L. Artif. Intell., 101 (1998) 285.
35. Bahler, D. and Bristol, D., Predictive Toxicology of Chemicals: Experience and Impact of AI tools (AAAI Spring Symposium Technical Report SS-99-01) AAAI Press Menlo Park CA, 1999, pp. 74–77.
36. Quinlan, J.R., C4.5: Programs for Empirical Learning, Morgan Kaufmann, San Fancisco, CA, 1993.