Predicting Chemical Carcinogenesis Using Structural Information Only*

Claire J. Kennedy[1], Christophe Giraud-Carrier[1], and Douglas W. Bristol[2]

[1] Department of Computer Science, Merchant Venturers Building, University of Bristol, Bristol BS8 1UB, U.K.
{kennedy,cgc}@cs.bris.ac.uk
[2] National Institute of Environmental Health Sciences, Box 12233, RTP, NC 27709 U.S.A.bristol@niehs.nih.gov

Abstract. This paper reports on the application of the Strongly Typed Evolutionary Programming System (STEPS) to the PTE2 challenge, which consists of predicting the carcinogenic activity of chemical compounds from their molecular structure and the outcomes of a number of laboratory analyses. Most contestants so far have relied heavily on results of short term toxicity (STT) assays. Using both types of information made available, most models incorporate attributes that make them strongly dependent on STT results. Although such models may prove to be accurate and informative, the use of toxicological information requires time cost and in some cases substantial utilisation of laboratory animals. If toxicological information only makes explicit, properties implicit in the molecular structure of chemicals, then provided a sufficiently expressive representation language, accurate solutions may be obtained from the structural information only. Such solutions may offer more tangible insight into the mechanistic paths and features that govern chemical toxicity as well as prediction based on virtual chemistry for the universe of compounds.

1 Introduction

This paper reports on the application of the Strongly Typed Evolutionary Programming System (STEPS) [4] to the *IJCAI* Predictive Toxicology Evaluation (PTE) challenge [8]. A second round of the challenge (PTE2) consists of predicting the outcome for 30 chemical bioassays for carcinogenesis being conducted by the National Institute of Environmental Health Sciences in the USA. The data provided includes both structural and non-structural information. The non-structural information consists of the outcomes of a number of laboratory analyses (e.g., Ashby alerts, Ames test results). The structural information is simply a graphical representation of the molecules in terms of atoms and bond connectives. Most contestants so far have relied heavily on results of short term toxicity (STT) assays. It appears that, for some learning tasks and systems, the addition of this type of information improves the predictive performance of the

* This work is funded by EPSRC grant GR/L21884

J.M. Żytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 360–365, 1999.

induced theories [10]. On the other hand, for other tasks and systems, the opposite seems to be true, i.e., propositional information has a negative effect on generalisation [3].

We strongly argue that, for the PTE2 task, if toxicological information only makes explicit, properties implicit in the molecular structure of chemicals, the non-structural information is actually superfluous. In addition, obtaining such properties often requires time cost and in some cases substantial utilisation of laboratory animals [1,6]. Provided a sufficiently expressive representation language, good solutions may be obtained from the structural information only. Hence, prediction can potentially be made faster and more economically. Experiments reported here with STEPS support this claim. The inherent structure of the graphical representation of molecules is captured naturally by the individuals-as-terms representation of STEPS. The rules obtained with structural information only are better in terms of accuracy than those obtained using both structural and non-structural information. In addition, our approach is more likely to provide insight into the mechanistic paths and features that govern chemical toxicity, since the solutions produced are readily interpretable as chemical structures. STEPS ranks joint 2nd of 10 in the current league table of the PTE2 challenge.

The paper is organised as follows. Section 2 describes the PTE2 challenge. Section 3 reports the results of applying STEPS to PTE2. Finally, section 4 concludes the paper.

2 The PTE2 Challenge

The National Institute of Environmental Health Sciences (NIEHS) in the USA provides access to a large database on the carcinogenicity (or non-carcinogenicity) of chemical compounds through the National Toxicology Program (NTP). The information has been obtained by carrying out long term bioassays that have classified over 300 substances to date. The Predictive Toxicology Evaluation (PTE) challenge was organised by the NTP to gain insight into the features that govern chemical carcinogenicity [2]. The first *International Joint Conference on Artificial Intelligence (IJCAI)* PTE challenge involved the prediction of 39 chemical compounds that were, at the time, undergoing testing by the NTP. The training set consisted of the remaining compounds in the NTP database. The participants consisted of both experts in the area of chemical toxicology and machine learning systems. Symbolic machine learning, and in particular Inductive Logic Programming, has been applied with great success to bio-molecular problems in the past [12,5]. Symbolic machine learning techniques are particularly suitable for problems of this type since it is not only the prediction that is interesting, but also the induced theory which provides an explanation for the predictions. The learning system Progol, for example, was entered into the PTE challenge and obtained results that were competitive with those obtained by the expert chemists [9].

Following on the success of the first challenge, a second round of the PTE challenge (PTE2) [7] was presented to the AI community at *IJCAI* in 1997 [11]. The PTE2 challenge involves the prediction of 30 new bioassays for carcinogenesis being conducted by the NTP. The training set consists of the remaining 337 bioassays in the NTP database. At the time of writing the results for 7 of the chemical compounds in the test set are still unknown. Ten machine learning entries have been made so far in reaction to the PTE2 challenge, and their performance has been calculated on the 23 chemical compounds whose results are known [8]. In addition to predictive accuracy, entries have been evaluated according to whether or not they exhibit explanatory power, where the explanatory power of a theory exists "... if some or all of it can be represented diagrammatically as chemical structures." [11].

The PTE challenges provide the machine learning/data mining communities with an independent forum in which intelligent data analysis programs and expert chemists can work together on a difficult scientific knowledge discovery problem.

3 Experiments

This section reports on experiments using STEPS on the PTE2 dataset. STEPS is a strongly-typed evolutionary system, which evolves program trees using constructs from the Escher programming language (see [4] for details).

3.1 Data

In order to tackle the PTE2 problem using STEPS the original Prolog representation [8], consisting of the 337 training cases, was translated into the Escher closed term representation.

Each chemical molecule is represented by highly structured term consisting of properties of the molecule and the atoms and bonds that form the structure of the molecule. The properties of the molecule resulting from laboratory analyses consist of Ames test results (i.e., whether the compound is mutagenic or not - mutagenicity is an indication of carcinogenicity), two sets of genetic toxicology test results, one for positive and one for negative results, and a set of Ashby alerts and their counts (properties of the molecule that are likely to indicate carcinogenicity, discovered by a toxicology expert).

The atom and bond structure that make up the molecule is represented as a graph, i.e., a set of atoms and a set of bonds connecting pairs of atoms. An atom consists of a label (which is used to reference that atom in a bond description), an element, one of 233 types represented by integers, and a partial charge. A bond is a tuple consisting of a pair of labels for the atoms that are connected by the bond and the type of the bond (e.g., single, double).

3.2 Method

The aim of PTE2 is to generate a concept description that can distinguish between *Active*, or carcinogenic compounds and *InActive*, or non-carcinogenic compounds. The concepts induced here are restricted to the following form:

```
IF Cond THEN C1 ELSE C2;
```

Since either *Active* or *Inactive* can be used for `C1` (leading to potentially different induced theories) and the experiments are intended to compare learning from structural-only information and learning from all available information, there are four settings to compare.

As STEPS is a stochastic algorithm the experiments are repeated ten times for each particular setting. The best performing theories as measured on the training data are output at the end of a run. The theory with the highest accuracy on the test set is then chosen as the best theory for that particular run. The best theory from the set of ten experiments is selected as the theory for a particular setting of data and description format.

The method of fitness evaluation used here is the Stepwise Adaption of Weights (SAW) method [13]. The SAW fitness function essentially implements a weighted predictive accuracy measure, which is based on the perceived difficulty of the examples to be classified. During evolution, only training examples are used. The SAW fitness function rewards an individual for the correct classification of a *difficult* example by associating a weight with each example. An example is considered difficult if the current best theory of the generation can not classify it correctly, in which case its associated weight is incremented by an amount `delta weight`. The weights are adjusted every `weight gen` generations. The fitness for a particular individual therefore becomes a weighted sum of the number of training examples that it can correctly classify.

The parameters for STEPS and SAW, for all experiments are as follows: delta weight = 0.1, weight gen = 5, population size = 100, maximum no. generations = 150, minimum depth = 3, maximum depth = 20, and selection = tournament.

3.3 Results

The following table gives the results for the ten runs for each of the four configurations. The Best Accuracy is the accuracy of the best performing theory out of all ten runs for a particular configuration.

Configuration (C1-Info)	Best Accuracy
Active-All	65%
Inactive-All	74%
Active-Struc	70%
Inactive-Struc	78%

The program selected as the best out of the forty runs with the various dataset and default class configurations achieved a predictive accuracy of 78%

on the test data and was obtained with the *Inactive-Struc* Configuration. This definition, which is currently undergoing more thorough analysis, is joint second in the PTE2 league table (see [8]). It is given here in Escher program format in Figure 1 and in its English equivalent in Figure 2.

```
carcinogenic(v1) =
       if
        ((((card (filter (\v3 -> ((proj2 v3) == O))
            (proj5 v1))) < 5) &&
          ((card (filter (\v5 -> ((proj2 v5) == 7))
            (proj6 v1))) > 19)) ||
        exists \v4 -> ((elem v4 (proj6 v1)) && ((proj2
           v4) == 3))) ||
        (exists \v2 -> ((elem v2 (proj5 v1)) &&
           ((((((proj3 v2) == 42) ||
           ((proj3 v2) == 8)) ||
           ((proj2 v2) == I)) ||
           ((proj2 v2) == F)) ||
           ((((proj4 v2) within (-0.812,-0.248)) &&
             ((proj4 v2) > -0.316)) ||
           (((proj3 v2) == 51) ||
           (((proj3 v2) == 93) &&
              ((proj4 v2) < -0.316))))))
        && ((card (filter (\v5 ->
                  ((proj2 v5) == 7))(proj6 v1))) < 15))
     then Inactive
     else Active;
```

Fig. 1. The best definition produced by STEPS as an Escher program

4 Conclusion

This paper reports on the application of STEPS, to the PTE2 challenge. The rules obtained by STEPS using structural information only, are comparable in terms of accuracy to those obtained using both structural and non-structural information by all PTE2 participants. In addition, this approach may produce insights into the underlying chemistry of carcinogenicity, one of the principal aims of the PTE2 challenge. Furthermore, as the theory produced by STEPS relies only on structural information, carcinogenic activity for a new chemical can be predicted without the need to obtain the non-structural information from laboratory bioassays. Hence, results may be expected in a more economical and timely fashion, while also reducing reliance on the use of laboratory animals.

References

1. D.R. Bahler and D.W. Bristol. The induction of rules for predicting chemical carcinogenesis in rodents. In *Intelligent Systems for Molecular Biology*, pages 29–37. AAAI/MIT Press, 1993.

```
A molecule is Inactive if it
            contains less than 5 oxygen atoms
               and has more than 19 aromatic bonds,
            or if it contains a triple bond
            or if it contains an atom that
                is of type 42 or 8 or 51
                or is an iodine or a fluorine atom
                or has a partial charge between -0.812 and -0.316
                or is of type 93 with a partial charge less than -0.316
             and contains less than 15 aromatic bonds
          Otherwise the molecule is active.
```

Fig. 2. The best definition produced by STEPS in English

2. D.W. Bristol, J.T. Wachsman, and A. Greenwell. The NIEHS predictive toxicology evaluation project. *Environmental Health Perspectives*, pages 1001–1010, 1996. Supplement 3.
3. P. Flach and N. Lachiche. 1BC: a first-order bayesian classifier. In *Proceedings of the Nineth International Conference on Inductive Logic Programming (ILP'99)*. LNCS, Springer, 1999.
4. C.J. Kennedy and C. Giraud-Carrier. An evolutionary approach to concept learning with structured data. In *Proceedings of the Fourth International Conference on Artificial Neural Networks and Genetic Algorithms*. Springer Verlag, 1999.
5. R. King, S. Muggleton, A. Srinivasan, and M. Sternberg. Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity in inductive logic programming. *Proceedings of the National Academy of Sciences*, 93:438–442, 1996.
6. Y. Lee, B.G. Buchanan, and H.R. Rosenkranz. Carcinogenicity predictions for a group of 30 chemicals undergoing rodent cancer bioassays based on rules derived from subchronic organ toxicities. *Environ Health Perspect*, 104(Suppl 5):1059–1064, 1996.
7. http://dir.niehs.nih.gov/dirlecm/pte2.htm.
8. http://www.comlab.ox.ac.uk/oucl/groups/machlearn/PTE/.
9. A. Srinivasan and R. King. Carcinogenisis predictions using ILP. In *Proceedings of the Seventh Inductive Logic Programming Workshop*. LNAI, Springer Verlag, 1997.
10. A. Srinivasan, R. King, and S. Muggleton. The role of background knowledge: Using a problem from chemistry to examine the performance of an ILP program. In *Intelligent Data Analysis in Medicine and Pharmacology*. Kluwer Academic Press, 1996.
11. A. Srinivasan, R.D. King, S.H. Muggleton, and M. Sternberg. The predictive toxicology evaluation challenge. In *Proceedings of the Fifteenth International Joint Conference Artificial Intelligence (IJCAI-97)*. Morgan-Kaufmann, 1997.
12. A. Srinivasan, S. Muggleton, R. King, and M. Sternberg. Mutagenesis: ILP experiments in a non-determinate biological domain. In *Proceedings of Fourth Inductive Logic Programming Workshop*. Gesellschaft für Mathematik und Datenverarbeitung MBH, 1994.
13. J.I. van Hemert and A.E. Eiben. Comparison of the SAW-ing evolutionary algorithm and the grouping genetic algorithm for graph coloring. Technical Report TR-97-14, Leiden University, 1997.