# Applying the Subdue Substructure Discovery System to the Chemical Toxicity Domain

**Ravindra N. Chittimoori, Lawrence B. Holder and Diane J. Cook**

Department of Computer Science Engineering
University of Texas at Arlington
Box 19015 (416 Yates St.), Arlington, TX 76019-0015
Email: {chittimo, holder, cook}@cse.uta.edu
Phone: (817)-272-2596
Fax: (817)-272-3784

**Abstract**

The ever-increasing number of chemical compounds added every year has not been accompanied by a similar growth in our ability to analyze and classify these compounds. The problem of prevention of cancer caused by many of these chemicals has been of great scientific and humanitarian value.

The use of AI discovery tools for predicting chemical toxicity is being investigated. The basic idea behind the work is to obtain structure-activity representation (SARs)[Srinivasan et al.], which relates molecular structures to cancerous activity. The data is obtained from the U.S National Toxicology Program conducted by the National Institute of Environmental Health Sciences (NIEHS). A general approach to automatically discover repetitive substructures from the datasets is outlined by this research. Relevant SARs are identified using the Subdue substructure discovery system that discovers commonly occurring substructures in a given set of compounds. The best substructure given by Subdue is used as a pattern indicative of cancerous activity.

## Introduction

The researcher's ability to interpret the data and discover interesting patterns within the data is of great importance as it helps in obtaining relevant SARs [Srinivasan et al.], for the cause of chemical cancers (e.g., Progol identified a primary amine group as a relevant SAR for the cause of the chemical cancers [Srinivasan et al. 1997]). One method for interpreting and discovering interesting patterns in the data is the identification of common substructures within the data. These substructures should be capable of compressing the data and identifying conceptually interesting substructures that enhance the interpretation of data. This identification also helps in simplifying the data by replacing instances of the substructure with a pointer to the newly discovered substructure. The subsequent iterations of the discovery and replacement process construct a hierarchical description of the structural data in terms of discovered substructures.

Discovering substructures for identifying relevant SARs has been a prominent area of application for knowledge discovery systems like the Subdue system [Cook and Holder 1994]. It discovers interesting substructures in structural data based on the Minimum Description Length [MDL] principle [Rissanen 1989]. Subdue discovers substructures that compress the original data and represent structural concepts in data.

This paper is organized as follows. The next section discusses the problem of chemical toxicity prediction. Then the mechanism by which the knowledge discovery system Subdue extracts molecular descriptions for attaining relevant SARs is explained in detail. The next section discusses the current domain (chemical toxicity). The methodologies used by the domain specialist to represent the data and the preliminary results are discussed in brief. The final sections talk about the conclusions and the future work in this area.

## Carcinogenesis Prediction Problem

The problem of prediction of carcinogenecity of a particular compound is of unquestionable importance. It is estimated that nearly 100,000 chemicals are in use in large amounts every day [Huff, Haseman and Rall 1991]. Many more chemicals are being added every year to the already existing set of chemicals. In light of this ever-growing increase in the number of chemicals, the U.S National Toxicology Program conducted chemical bioassays to help in identifying substances that may have a carcinogenic effect. The process of obtaining empirical evidence from such rodent bioassays is too expensive and too slow to cope with the ever-increasing number of chemicals. (On an average each rodent is exposed to a chemical for a period of two years and an average of 500-1000 chemicals are added every year). Hence an urgent need for models that propose molecular mechanisms for carcinogenesis is envisaged. It is believed that these models would cut down the costs, reduce dependence on laboratory animals and generate reliable toxicity predictions for all kinds of chemicals.

## Overview of SUBDUE

The Subdue system discovers the substructures in the databases that compress the original data and represent structural concepts in the data. The best substructure is found after multiple passes by replacing the previously discovered substructures in each pass. A substructure is a connected subgraph within the graphical representation. The discovery system represents structural data as a labeled graph. Objects in the data map to vertices or small subgraphs in the graph, and relationships between objects map to directed or undirected edges in the graph. This graphical representation serves as input to Subdue (e.g., see figure 1). The discovery algorithm used by Subdue is a computationally constrained beam search. The algorithm begins with the substructure matching a single vertex in the graph. The algorithm selects the best substructure in each iteration and incrementally expands the instances of a substructure. An instance of a substructure in an input graph is a set of vertices and edges from the input graph that match, graph theoretically, to the graphical representation of the substructure. These new substructures become candidates for further expansion. This algorithm searches for the best substructure until all possible substructures have been considered or the total amount of computation exceeds a given limit. Evaluation of each substructure is determined by how well the substructure compresses the description length of the concerned database.

To identify substructures that occur often in data but not always in the same form, Subdue uses a computationally bounded inexact graph match [Bunke and Allerman 1983]. The inexact substructure discovery can be used to discover interesting structures in the input data, whose instances are found either in the same form or in a slightly convoluted form. Subdue's search can be guided towards appropriate substructures for a particular domain (in our case the chemical toxicity domain) by the inclusion of background knowledge (e.g., known relevant SARs).

## Chemical Toxicity Domain

A database of more than 300 chemicals has been created due to the tests conducted by the U.S. National Toxicology Program (NTP) [niehs]. These compounds are determined to be carcinogenic or noncarcinogenic. Levels of evidence of carcinogenecity are obtained from the incidence of tumors on long term exposure to chemicals using rats and mouse strains as predictive surrogates for humans. The NTP assigned the following levels of evidence for the compounds: CE – clear evidence of cancerous activity, SE – some evidence of cancerous activity, E – equivocal evidence of cancerous activity and NE – no evidence of cancerous activity. Conventional regression based techniques ([Kubini 1993]) cannot be applied to model the compounds in the NTP database because of the diversity of the compounds present. Hence the need for some discovery algorithms that can discover interesting, useful concepts even in the most varying domains.

The datasets in the NTP database contain information about more than 300 chemical compounds that are either carcinogenic or noncarcinogenic. Primarily there are 298 chemical compounds whose carcinogenecity is known. This comprises the training set of the Subdue program. There are 69 compounds whose carcinogenecity is not known. This comprises the experimental set of the Subdue program. The information in these sets relates to the molecular structures of the compounds, and includes the atoms, bonds and domain specific knowledge about various groups like alcohol, amine, amino, benzene, ester, ether, ketone, methanol, methyl, nitro, phenol and sulfide. The representation also contains information about the compound test results (+/-) on the various properties of carcinogenecity like Ames test, Chromex, Chromaberr, Drosophilia, Mouse-Lymph, Salmonella Assay. The aim of this research project is to obtain SARs despite the diversity present among the compounds.

## Methodology

The training set is further divided into positive examples and negative examples. Subdue is applied to the positive (cancerous) and the negative (non-cancerous) examples separately and the best substructures are identified in each of these training sets. The resultant best substructures from each of the two training sets (positive and negative) are compared. The substructures that occur in the positive examples but not in the negative examples are identified. These identified substructures are used as the pattern indicative of cancerous activity.

The toxicity of the chemicals in the experimental set can be determined by the following approaches. One approach is to apply Subdue individually to the compounds in the experimental set and record the best substructure in each of the compounds. Based on the judgement of the domain specialist (comparing the best substructure returned by Subdue with the substructures identified from the training set) the compound in the experimental set is determined to be carcinogenic or noncarcinogenic. Presently we are using this approach in identifying the carcinogenicity of compounds in the experimental set. The second approach is to include substructures identified in the training set as predefined substructures for Subdue in its search on the experimental set. Subdue will first search the input graph of the compound for instances of the predefined substructures, using inexact graph matches. Instances that match within the inexact match threshold are subsequently expanded. The domain specialist determines the carcinogenecity or noncarcinogenecity of a compound in the experimental set depending on how well the predefined substructure helped in compressing the description length of the compound. The third approach is to check if the discovered substructure SAR appears anywhere in the compound to be classified. Once unique SARs are discovered, the presence of only one substructure might be enough evidence to predict carcinogenecity.

The input to the Subdue program is the graphical representation of all the chemical compounds. Each of the atoms in a compound is represented as a vertex and the bonds between the atoms are represented as undirected edges between the vertices. Domain knowledge is incorporated into Subdue to guide the discovery process. Various groups like methyl, benzene, amino etc., each are represented as a vertex in each compound and have directed edges to all the atoms in a compound, which participate in the group. Properties like Salmonella assay and Ames test are each represented as a vertex in each compound and have directed edges to all the atoms in the compound with the string label on each of these edges specifying whether the compound tested positive or negative on this property. Figure 1 shows a sample graph. To capture the diversity present in the atoms (atom name, atom type, and partial charge), each of the atoms is represented as a separate node with directed edges to the name of the atom (n), type (t) and partial charge (p). The relationship between atoms (i.e. bonds) is represented as undirected edges.

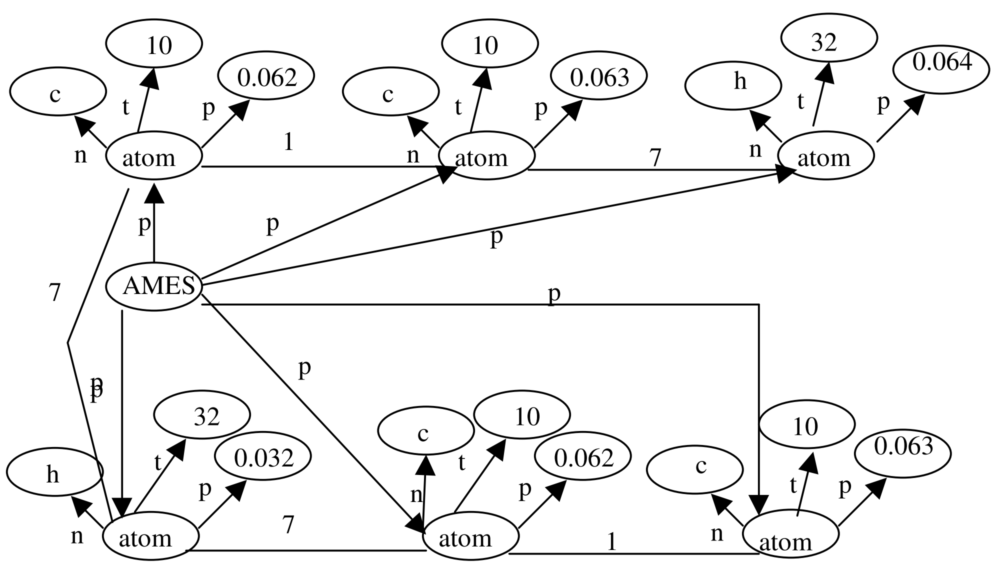


**Figure 1: Results of Subdue on part of chemical compound.**

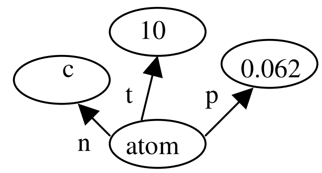


**Figure 2: Substructure $S_1$ discovered by Subdue.**

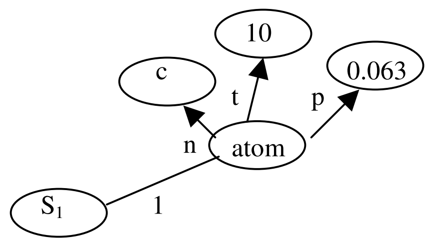


**Figure 3: Substructure $S_2$ discovered by Subdue.**

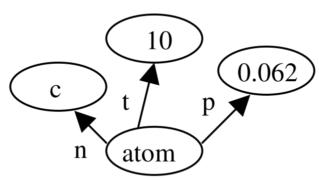


**Figure 4: Substructure with 4 vertices discovered by Subdue.**

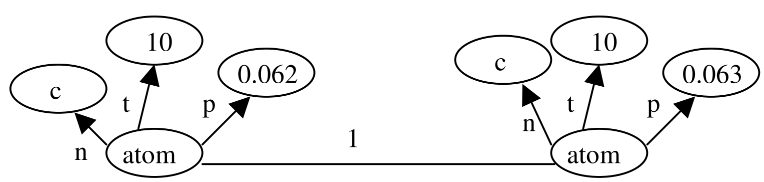


**Figure 5: Substructure with 8 vertices discovered by Subdue.**

Subdue discovers substructure $S_1$ in the compound as in figure 2 above. $S_1$ when used to compress the sample graph further, finds substructure $S_2$ as in figure 3. Subdue generates a similar hierarchical description of structures with such repeated applications.

## Results

Subdue has been successful in discovering small substructures. Figure 4 is an example of a substructure discovered. Subdue was successful in discovering the exact number of instances of this substructure in the positive (cancerous) examples. Figure 5 is an improvement over figure 4. Subdue discovered a substructure with 8 vertices. Efforts are being made in guiding Subdue to discover more complex substructures that might help in relating a compound with carcinogenecity.

The parameter settings for Subdue help in guiding the search towards a more specific result. The parameters that might affect the results of Subdue most are the threshold parameter (the fraction of the size of an instance by which the instance can be different) and the size parameter (size of considered substructures). By specifying a value for the threshold parameter, the inexact graph match can be done. If the domain specialist believes that the partial charge value need not be exactly matched, but can vary by a range of (+/-) x, then Subdue can be tuned to perform the inexact graph match by specifying a value for the threshold parameter. The size of the substructure considered by Subdue can be specified by the size parameter that has a lower bound and an upper bound. If the domain specialist believes that the substructure discovered by Subdue is too inconsequential or too big to find any relevant SARs, then appropriate values can be specified to guide Subdue in its discovery. We are optimistic that relevant SARs that indicate carcinogenic activity can be identified by Subdue.

## Conclusions

The prediction of carcinogenecity and the modeling of diverse chemical compounds is of unquestionable importance. The data mining algorithms capable of handling the increasing structural component of today's databases can achieve this. Subdue, a data mining algorithm, is specifically designed to discover increasingly interesting and repetitive patterns within the data that relates molecular structure to cancerous activity.

In this paper, the methodologies of representing the chemical toxicity domain are discussed at length. The initial results of Subdue are explained and an effort is made to explain the eventual capability of Subdue to discover a pattern that distinguishes carcinogenic and noncarcinogenic compounds.

## Future Work

Future research aims at describing the possible relationships between molecular structure of a compound on the one hand, and biological and toxicological processes on the other. Subdue is presently in an experimental phase. Making use of parallel and distributed resources can significantly improve the run-time performance of data-intensive and compute-intensive discovery programs such as Subdue. We are currently evaluating the benefits of applying a parallel version [Galal, Cook and Holder 1997] of Subdue on the chemical toxicity domain.

## References

[Srinivasan et. al] A. Srinivasan, S.H. Muggleton, R.D. King and M.J.E. Sternberg. The Predictive Toxicology Evaluation Challenge. http://www.comlab.ox.ac.uk/oucl/groups/machlearn/PTE/

[Cook and Holder 1994] D.J.Cook and L.B.Holder. Substructure Discovery Using Minimum Description Length and Background Knowledge. In *Journal of* Artificial *Intelligence Research,* Volume 1, pages 231-255, 1994.

[Rissanen 1989] J.Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific Publishing Company, 1989.

[Huff, Haseman and Rall 1991] J.E.Huff, J.K.Haseman, and D.P.Rall. Scientific concepts, value and significance of chemical carcinogenesis studies. *Ann Rev Pharmacol Toxicol,* 31:621-652, 1991.

[Bunke and Allerman 1983] H.Bunke and G.Allermann. Inexact graph matching for structural pattern recognition. *Pattern Recognition Letters,* 1(4):245-253, 1983.

[niehs] http://ntp-server.niehs.gov/main_pages/about_NTP.html

[Kubini 1993] H.Kubini. *QSAR: Hansch Analysis and Related Approaches.* VCH, New York, 1993.

[Galal, Cook and Holder 1997] G.Galal, D.J.Cook and L.B.Holder, Improving Scalability in a Knowledge Discovery System by Exploiting Parallelism. In the *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining,* pages 171-174. 1997.