Josef Kittler Fabio Roli (Eds.)

Multiple Classifier Systems

Second International Workshop, MCS 2001
Cambridge, UK, July 2-4, 2001
Proceedings

Springer

Series Editors

Gerhard Goos, Karlsruhe University, Germany
Juris Hartmanis, Cornell University, NY, USA
Jan van Leeuwen, Utrecht University, The Netherlands

Volume Editors

Josef Kittler
University of Surrey, Centre for Vision, Speech and Signal Processing
Guildford, Surrey GU2 7XH, UK
E-mail: j.kittler@eim.surrey.ac.uk

Fabio Roli
University of Cagliari, Department of Electrical and Electronic Engineering
Piazza d'Armi, 09123 Cagliari, Italy
E-mail: roli@diee.unica.it

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Multiple classifier systems : second international workshop ; proceedings / MCS 2001, Cambridge, UK, July 2 - 4, 2001. Josef Kittler ; Fabio Roli (ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ; Milan ; Paris ; Singapore ; Tokyo : Springer, 2001
(Lecture notes in computer science ; Vol. 2096)
ISBN 3-540-42284-6

CR Subject Classification (1998): I.5, I.4, I.2.10, I.2, F.1

ISSN 0302-9743
ISBN 3-540-42284-6 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Printed in Germany

Typesetting: Camera-ready by author, data conversion by PTP-Berlin, Stefan Sossna
Printed on acid-free paper SPIN: 10839435 06/3142 5 4 3 2 1 0

Mixing a Symbolic and a Subsymbolic Expert to Improve Carcinogenicity Prediction of Aromatic Compounds

Giuseppina Gini,[1] Marco Lorenzini[1], Emilio Benfenati[2], Raffaella Brambilla[2], and Luca Malvé[2]

[1] Dept of Electronics and Information, Politecnico di Milano, piazza L. da Vinci 3", 20133 Milano, Italy
gini@elet.polimi.it
[2] Dept. of Environmental Health Sciences, Istituto di Ricerche Famacologiche "Mario Negri", Via Eritrea 62, 20157 Milano, Italy Milano, Italy
benfenati@marionegri.it

Abstract. One approach to deal with real complex systems is to use two or more techniques in order to combine their different strengths and overcome each other's weakness to generate hybrid solutions. In this project we pointed out the needs of an improved system in toxicology prediction. An architecture able to satisfy these needs has been developed. The main tools we integrated are rules and ANN. We defined chemical structures of fragments responsible for carcinogenicity according to human experts. After them we developed specialized rules to recognize these fragments into a given chemical and to assess their toxicity. In practice the rule-based expert associates a category to each fragment found, then a category to the molecule. Furthermore, we developed an ANN-based expert that uses molecular descriptors in input and predicts carcinogenicity as a numerical value. Finally we added a classifier program to combine the results obtained from the two previous experts into a single predictive class of carcinogenicity to man.

1 Introduction

The goal to predict carcinogenicity is a challenging one, in consideration of the social and economical importance of the problem. Chemicals are responsible for many tumors, and industry is required to take into account carcinogenicity of the chemicals used and produced. However, the experimental tests on chemicals last for years, are costly and require the use of animals, with the consequent ethical problems. Considering the importance of the goal, it is interesting to continue the attempts to improve computerized systems to predict carcinogenicity. So far the most popular programs have been expert systems [1] (ES). In many cases they look for the presence of toxic residues in the molecule, as in [2]. More recently neural networks (ANN) have been used [3], and inductive learning [4].
In the present study we tried a new approach, combining different systems in hybrid architecture. We developed an ES able to recognize toxic residues predicting a class of toxicity. Furthermore, we used ANN with molecular descriptors as input to provide

J. Kittler and F. Roli (Eds.): MCS 2001, LNCS 2096, pp. 126–135, 2001.

a different prediction of toxicity. Finally, we used a symbolic rule induction program to merge the information from the two sources.

2 Definition of the Phenomenon to Be Modeled: Carcinogenicity

Cancer is not a single disease. Furthermore, each single cancer involves a complex sequence of events. The complexity of the phenomenon means that experimental data are not precise, and in some cases contradictory. Most of the experiments are done on animals. Extrapolation of results from animals to humans is complicated also because in animal experiments high doses are used, while humans are generally exposed to low doses.

Carcinogens are listed in classes by national and international agencies. The International Agency on Research on Cancer (IARC) considers four classes: the compounds which have been recognized as carcinogenic to man are in *class 1*, the compounds which are not carcinogenic (only a few compounds) are in *class 4*, the other compounds are split in three classes of different degree of uncertainty: probably or possibly carcinogenic to man (*class 2A and 2B*), not classifiable as their carcinogenicity to humans (*class 3* - the most numerous one, characterized by the highest uncertainty). This classification combines, in the evaluation of carcinogenicity, the experimental evidences with the amount of epidemiological knowledge available.

A different approach has been introduced by Gold and colleagues [5]. They developed a numerical data set that contains standardized and reviewed results for carcinogenicity for more than 1200 chemicals. The cancerogenicity data on rat and mouse are expressed in term of the parameter TD50, which is the chronic dose rate, which would give half of the animals tumors within some standard experiment time. The huge amount of data and the quantitative homogeneous evaluation represented two important advantages.

Both kinds of characterization have been used: categorical (as the IARC) and continuous (as the Gold data set), the first with the residue approach, the second with the ANN. We extended its applicability to man using a symbolic rule induction program. To do this, for the training of this module we used the IARC classification.

In the area of toxicity prediction QSAR (Quantitative Structure Activity Relationships) and SAR (Structure Activity Relationships) models are common. They are based on the evidence that the structure of a molecule is responsible of its activity, and that biological data about the mechanisms are not a must to predict the outcome. Generally SAR models for carcinogenicity are only able to classify in two classes: positive or negative, while QSAR models give a real value for the toxicity.

Usually QSAR are methods to assess drugs; the challenge is to use them to predict toxicity values for large classes of chemicals and for complex phenomena as cancer.

3 Our Residue Approach

Many toxicologists consider the presence of given fragments in the molecule as an indication of potential carcinogenicity.

For instance, Ashby and Paton [6] listed many toxic residues responsible for adverse activity. CompuDrug at the end of the eighties started from this approach and encoded into its ES, called HazardExpert, the behaviour of selected residues based on a report by the U.S. Environmental Protection Agency. To enhance the efficiency of the system, there are built-in modules, which predict the dissociation constant (p*Ka*) and the distribution coefficient (log*P*). These can be used to predict the bioabsorption and accumulation of xenobiotics in living organisms, which have already been predicted, in addition to oncogenicity, mutagenicity, teratogenicity, irritation, sensitization, immunotoxicity and neurotoxicity. HazardExpert examines the compound itself as well as potential metabolites, based on modules providing for generation of potential metabolites. We made an accuracy test on HazardExpert in 1995 in the European project EST, and we found ways to improve it [7].
Sanderson and Earnshaw [8] used the rules *if..then* introducing a series of substructures known to be toxic in the rule base of a system called DEREK (Deductive Estimation of Risk from Existing Knowledge), that then recognizes any such residues in the compound examined. DEREK makes qualitative rather than quantitative predictions. It looks for previously characterized toxicophores that are highlighted in the display and their toxic activity associated. The presence of several toxicophores in the molecule means there are more risks, but whether the risks are additive or not is decided by the user. also DEREK takes into account physicochemical properties such as log*P* and pKa. There are several toxicological endpoints including mutagenicity, carcinogenicity, skin sensitization, irritation, reproductive effects, neurotoxicity and others.

3.1 Definitions of Rules of Ar-N Compounds

We studied this topic for aromatic amines and related compounds; in particular, we considered all the aromatic compounds with at least a nitrogen linked to the aromatic ring (Ar-N compounds), that contain a large number of chemicals, many of them carcinogens. The Ar-N group is divisible into 10 chemical classes further split into some subclasses, as shown in Table 1.

While classes are defined only considering the presence of a chemical group characterizing the Ar-N bond, subclasses are bounded by the following criteria:

1. presence of the same atom or substituent or chemical structure in a fixed position relative to the Ar-N bond,
2. implementation convenience: in order to reduce memory needs and accelerate the computer search;
3. toxicological affinity of chemicals in terms of TD50 values, target tissue and/or IARC class.

The structure for implementing this knowledge is a two-level structure, as illustrated in Fig. 1.

Table 1. Ar-N compounds divided into classes and subclasses.

1) PRIMARY AMINES
a- Monocyclic aromatic primary amines
b- Pentaatomic heteroaromatic primary amines
c- Hexaatomic heteroaromatic primary amines
d- Biphenyl primary amines
e- Di- and triphenylmethane amines and analogues
f- 4- and 4,4'-Stilbenes
g- 2-aminofluorene and analogues
h- Condensed polycyclic primary aromatic amines 1
i- Condensed polycyclic primary aromatic amines 2
2) NITROCOMPOUNDS
a- Monocyclic aromatic nitro compounds
b- 2-nitro-5-furyl
c- Thio- and azo-pentaatomic nitro compounds
d- Condensed polycyclic nitro compounds 1
e- Condensed polycyclic nitro compounds 2
f- Miscellaneous nitro compounds
3) AZOCOMPOUNDS
a- Dibenzo azo compounds
b- 1-naphtho azo compounds
c- 2-naphtho azo compounds
4) HYDRAZINES
a- Hydrazines 1
b- Hydrazines 2
5) SECONDARY AMINES
a- Aromatic secondary aliphatic amines
b- Diphenyl secondary amines
c- Carbazole
d- Solfonic secondary amines
e- Purines
6) AMIDES
a- Monocyclic aromatic amides
b- Biphenyl amides
c- 2-acetylaminofluorene derivatives
d- Pentaatomic heteroaromatic amides
e- Hexaatomic heteroaromatic amides
7) TERTIARY AMINES
a- Monocyclic aromatic tertiary amines
b- Di- and triphenylmethane tertiary amines
c- N,N-dihydroxyethyl tertiary amines
d- Nitrogen mustards
e- Pentaatomic heterocyclic tertiary amines
8) C-NITROSOCOMPOUNDS
9) N-NITROSOCOMPOUNDS
10) ISOCYANATES

For each subclass a first level structure, which identifies the chemical fragment common to each residue belonging to the subclass, has been individuated. The second level structures specify each residue. Two corresponding inhibition levels have been introduced for situations where the found fragment has no effect.

- *First level*: identifies the structure of the nitrogen fragment characterizing the class and the aromatics structures bonded to that group.
- *First inhibition level*: it solves the problem of compounds that, even if related to the structure of the subclass, are not carcinogens or have been ascribed to another subclass.
- *Second level*: the second search level permits the identification of a specific compound or small groups of compounds that refer to the same subclass but differ for some specific elements bound to the nitrogen group and/or to the aromatic structure, and suspected to be involved in the carcinogenicity process.
- *Second inhibition level*: this second inhibition level is useful to exclude a specific compound or a small group of compounds.

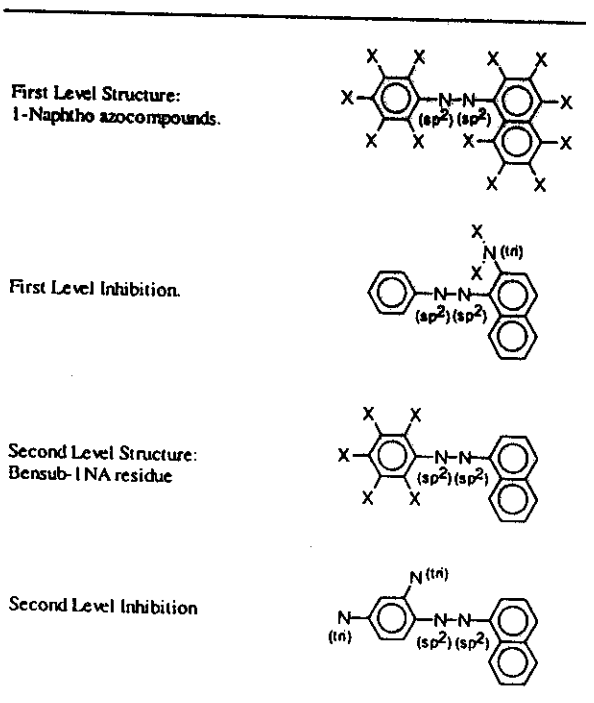

Fig. 1. Example of structure levels. The Figure shows the structure to search at the first and second levels and the relevant inhibitions.

Each fragment is associated with a category expressing the level of toxicity. Our system reports the highest level obtained and the residue responsible; if more than a

toxic residue is present, the program selects the most active. We defined five "carcinogenicity levels", using three parameters:
- the TD50 of the molecules [5];
- the level of carcinogenicity ascribed to the fragment contained in the molecule (averaging the evaluation for each fragment on all molecules containing the substructure);
- the classification or the evidence of carcinogenicity given by the databases IARC, IRIS, HSDB, NTP, RTECS.

The COSMIC format has been chosen to describe the molecules; it uses atom hybridization instead of information on atomic bonds, with two positive consequences:

- All bonds are equals. The chemical information is hidden in the nodes and so the search algorithm is easier.
- Hydrogens are left out. The molecular graph is smaller and so the search is faster.

3.2. Internal Representation

Graph theory was used to represent the chemical structures. They are stored as adjacency lists: given the *node i*, the nodes in the list *l* contain atoms that are adjacent to vertex *i* as shown in Figure 2.

3.3 Search Method

The search of a fragment in a molecule is a *subgraph isomorphism problem.* A graph G_α is *isomorphic* to a subgraph of a graph G_β if and only if there is a one to one correspondence between the node sets of this subgraph and those of G_α that preserves adjacency. The computational complexity of this problem is, in general, NP-Complete [9]. The search operation has been divided into two parts: the first search level is performed by finding all possible isomorphisms between the structure considered and the molecule, with the Ullmann's algorithm [9], modified to manage hydrogens and wildcards. After finding a first level structure, the second part of the search procedure checks positive and negative conditions, using a backtracking technique. If a second level structure and no inhibition are found, a residue is considered found.

4 The ANN-Based Prediction

Backpropagation neural network [10] has been adopted in this study to implement the quantitative prediction of carcinogenicity; more details are in [11].
From the Gold's database 104 molecules presenting an aromatic ring and a nitrogen linked to the aromatic ring have been chosen. We computed molecular descriptors of six main groups (physico-chemical, geometric, topological, electrostatic, quantum-chemical and thermodynamic); from the initial set of 34 descriptors a selection was necessary in order to avoid an excessive time for training the network. Principal

component analysis (PCA) has been used for the selection, building a final set of 13 descriptors (molecular weight, HOMO, LUMO, dipole moment, polarizability, Balaban, ChiV3 and flexibility indices, logD at pH 2 and pH 10, third principal axis of inertia, ellipsoidal volume, electrotopological sum).

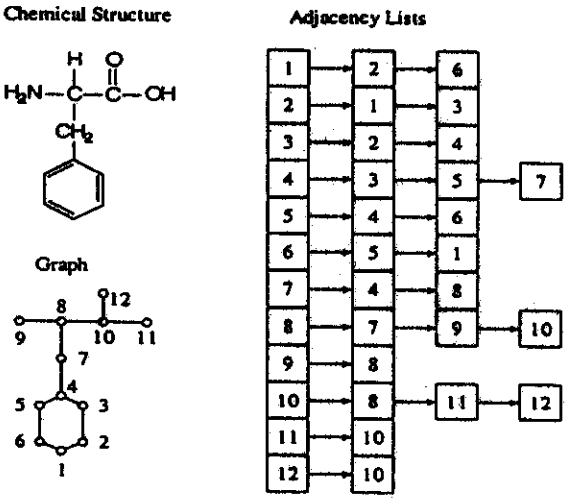

Fig. 2 The implementation through the adjacency lists.

The parameter TD50 reported by Gold et al. [11] has been adopted for the output. The output has been derived from a transformation of the TD50 according to the following formula:

$$\text{output} = \text{Log}\,(MW*1000/TD50)$$

Data were scaled between 0 and 1. The output has been also scaled accordingly. The scaling was based on the training set. Validation set was scaled on the basis of scaling of the training set.

All simulations were performed using MBP v 1.1 [10], initialing the weight with the SCAWI technique, and using the acceleration factors. Each network has been trained starting from 100 points random in the space, in order to minimize the probability of converging towards local minima.

For validation the N/2-fold-crossvalidation has been used. MSE and R^2_{cv} resulting from 10000 iterations of the back-propagation ANN, using different numbers of internal neurons, showed best results using four or seven hidden neurons: R^2_{cv} was in both cases 0.691.

The presence of outliers in the set has been supposed and investigated; 12 molecules were identified as outliers and removed. Results after outliers removal showed clear improvement in the R^2_{cv} which became 0.824 (with 4 hidden neurons). The majority (9 out of 12) of the outliers is molecules for which the experimental results were not statistically significant and an arbitrary 10^{31} value was given by Gold.

5 Combining the Two Predictions into the Hybrid System

The results we obtained from the two parts of the prediction should now be combined. In the present study we added a third module dedicated to the classification. Given the output of the residues research, and the expected TD50, we wanted to extrapolate a combined prediction of the human carcinogenicity.
We split some classes of the IARC classification according to the following criteria:
- to define 5 classes, 1 to 5, from lower to higher risks, based on the TD50 values;
- to check the presence of each residue in the molecules under study;
- to give to each residue a toxicity class obtained as the mean of the toxicity of the molecules where it was found;
- to assign to the molecule the maximum toxicity obtained from the residues and ANN module.

We built classification trees from examples, using different tree construction programs:

- C4.5 [12] which makes use of the maximization of the entropy gain, and builds hyper-rectangular in the attribute space;
- CART, which builds binary trees [13];
- OC1 [14], that uses a random perturbation of parameters to escape from local minima.

The training set has been prepared with all molecules and two attributes each, TD50 (predicted by BNN) and the carcinogenicity category predicted by the residue module. Performances using the leave-one-out are in Table 2.

Table 2. Results obtained with tree induction systems (accuracy %)

	C4.5	CART	OC1
Training	93.3	88.5	90.2
Validation	81.9	85.5	82.8

6 Discussion and Conclusions

Results in Table 2 show the accuracy (the ratio between the sum of correct assignments and the total compounds) show promising possibilities. The integration of the two approaches improved the performances of the individual methods.
Very few comparisons have been made of different ES in toxicology. Most of the papers presented by the authors of the different ES claim good predictions, often better than 90%. Omenn [15] reported the results of predictions on 44 chemicals made by some human experts and different computer programs. Table 3 compares the results with the ES and the best human results. This indicates that for the time being no ES can do better than a good human expert.

A particular problem is the nature and evaluation of the information. In several cases experts pay special attention to some data and overlook others, because they know from experience which data are most reliable. Sometimes their experience is

concentrated on certain aspects of the problem. As a consequence, different experts will give different answers.

Table 3.- ES and human experts predictions for toxicology of 44 chemicals

Expert	Accuracy	Percentage
Human experts	30/40	75
DEREK	22/37	59
TOPKAT	14/24	58
COMPACT	19/35	54
CASE	17/35	49

For this reason to assess the reliability of predictive models we must rely on internal evaluation, mainly on leave-one-out. The best we can expect is to be able to correctly predict a new external set, as we will try in the future.

In the attempt to overcome the limitation of attribute-based learning, some programs learn first-order predicate logic. Given background knowledge (expressed as predicates, positive examples, and negative examples) the ILP system is able to construct a predicate logic formula H such as all the positive examples can be logically derived from the background formulas and H, and no negative example can be logically derived.

The main work in ILP is the predictive toxicology challenge, which aims at constructing a SAR model based on data from NTP (National Toxicology Program). The NTP produced the PTE data sets, based on the study of about 300 compounds, to be used as training set, and the definition of small tests sets (30 compounds). For all molecules the carcinogenicity is available, expressed as yes/not. The data set presents a mix of chemicals (both organic and inorganic as representative of 19 millions); some chemical classes are not represented, some known biological mechanism is not represented.

In [16] a report of the submissions to the challenge is shown. In the models, presented by 9 laboratories, the best estimated accuracy is 0.87 for a stocastic system, on the outcome of 23 of the 30 molecules. The other models range from 0.78 o 0.48. The approach based on ILP reached 0.78.

Our research confirms the feasibility of an ANN for carcinogenicity for several chemical classes, which exhibit their activity according to different mechanisms. A valuable characteristic of our ANN is that it seems to correctly predict carcinogenic compounds; unfortunately, it is less accurate in the prediction of non-active compounds. It is likely that ANN alone cannot solve all the problems linked with carcinogenicity prediction. A classical example is the case of ortho- and para-anisidine that have very similar descriptors values, but one of the compounds is carcinogenic, while the other not. In this case an approach based on the residues can distinguish the two chemicals.

An advantage of our architecture, which evolved from a previous work on fitotoxicity [17] is that the output is not simply a classification into two classes of activity: carcinogenic or not, as in several programs predicting toxicity. Our system gives a quantitative prediction of the activity, and also a classification similar to IARC.

Moreover, we do bot need biological data to predict carcinogenicity. A key advantage of programs based on the simple chemical structure is that they do not require the synthesis of the chemical to be tested and biological experiments in order to make prediction. However, in the hybrid architecture we defined it is easy to introduce in the rule induction program other inputs, such as results from mutagenicity tests.

Acknowledgements. The European Union contracts COMET and IMAGETOX.

References

1. Benfenati, E., and Gini, G.:Computational predictive programs (expert systems) in toxicology, Toxicology, vol.119 (1997).213-225,.
2. HazardExpert, version 3.0. Compudrug Chemistry Ltd, Budapest, Hungary.
3. Benigni R. and Richard, A.M.: QSARS of mutagens and carcinogens: two case studies illustrating problems in the construction of models for noncongeneric chemicals, Mutation Res., vol.371 (1996).29-46.
4. Lee, Y., Buchanan, B. G., Mattison, D. M., Klopman, G., and Rosenkranz, H. S.: Learning rules to predict rodent carcinogenicity of non-genotoxic chemicals. Mutation Res., vol. 328 (1995) 127-149.
5. URL: http://sciweb.lib.umn.edu/s&e/chem.htm.
6. Ashby, J., and Paton, D.: The influence of chemical structure on the extent and sites of carcinogenesis for 522 rodent carcinogens and 55 different human carcinogens exposures. Mutation Res. 286 (1993) 3-74.
7. Darvas, F., Papp, A., Allerdyce, A., Benfenati, E., Tichy, M., Sobb, N., Citti, A., Gini, G.: Overview of Different Artificial Intelligence Approaches Combined with a Deductive Logic-based Expert System for Predicting Chemical Toxicity. In Gini and Katritzky (eds) AAAI Spring Symposium on Predictive Toxicology, SS-99-01, AAAI Press, Menlo Park, California (1999) 94-99.
8. Sanderson, D.M., and Earnshaw, C.G.: Computer prediction of possible toxic action from chemical structure; the DEREK system. Human and Experimental Toxicology 10 (1991) 261-273.
9. Ullmann, J. R.: An algorithm for subgraph isomorphisms. Journal of ACM, Vol. 23 - 1 (1976) 31-42.
10. Anguita, D.: Matrix Back Propagation v. 1.1 User Manual, Genova, Italy (1993).
11. Gini, G., Benfenati, E., Lorenzini, M., Bruschi, M., Grasso, P.: Predictive Carcinogenicity: a Model for Aromatic Compounds, with Nitrogen-containing Substituents, Based on Molecular Descriptors Using an Artificial Neural Network. J of Chem Inf and Comp Sciences, 39 (1999) 1076-1080.
12. Quinlan, J.R.: C4.5 Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Mateo, CA (1993).
13. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J.: Classification and regression trees. Wadsworth: Belmont, CA (1984).
14. Murthy, S. K., Kasif, S. and Salzberg, S.: A system for induction of obliques decision trees. J. of Artificial Intelligence Research, vol.2 (1994).
15. Omenn, G.S.: Assessing the risk assessment paradigm. Toxicology 102 (1995) 23-28.
16. Srinivasan, A., King, R. D., Bristol, D. W.: An assessment of submissions made to the predictive toxicology evaluation challenge. Proc.IJCAI 1999 (1999) 270-275.
17. Gini, G., Benfenati, E., Testaguzza, V., Todeschini, R.: Hytex (Hybrid Toxicology Expert System): Architecture and implementation of a multi-domain hybrid expert system for toxicology. Chemometrics and intelligent laboratory systems 43 (1998) 135-145.