Structure-activity relationship suggestion by a neuro-fuzzy logic engine
Jonathan Vessey1 and Elizabeth Colbourn2

INTRODUCTION

The discovery of structure-activity relationships (SARs) is a crucial step in the production of many prediction systems. In Lhasa Limited's Derek toxicity prediction system, SARs are the basis of toxicity alerts expressed in part as structural alerts, which describe toxicophores or mechanistic cause-and-effect relationships discovered by human scientists. Whist this results in a high quality, transparent knowledge base, the development of new knowledge is time consuming.

The suggestion of SARs by machine learning techniques is therefore potentially a valuable tool in facilitating the knowledge development of systems like Derek.

We have investigated FormRules, a program with a neuro-fuzzy logic engine, to suggest SARs suitable for further refinement by human knowledge developers.

About FormRules

FormRules is a desktop Windows application developed by Intelligensys Ltd., which has been designed to generate models – expressed as rules – about compound formulations. The rules take the form of if...then... statements, alongside measures of support for each statement in the training data set.

A model of any particular property commonly takes the form of several sub-models which themselves contain the rules. In generating the model – and discovering the rules – the user can limit the number of sub-models developed and the extent and method by which a model is refined.

Rules can contain many different properties linked by Boolean AND statements; below is an example FormRules model. Version 3.32 of FormRules was used in this study.

Models with Dragon descriptors

FormRules uses a neuro-fuzzy logic engine to discover rules and build models. We can utilise this to create rules from any data set. Using a mutagenicity data set, we generated models relating activity to descriptors obtained from Dragon2, using either subsets or all of the data, and varying the FormRules parameters. Below is an example model built using this approach showing typical results: the model has been built from subset CV1 of the data set (982 compounds) and 140 possible descriptors, identified below using their Dragon abbreviations.

Properties can have numeric or text values

Models produced using these descriptors are difficult to extract new Derek alerts from. Many known toxicophores already included in Derek are not captured well by the descriptors and so do not appear in the generated rules.

We find that predefined lists of fragment-based descriptors are ineffective in identifying key known toxicophores; even use of the 27000 fragment descriptors in Leadscope Enterprise is unsuccessful! Therefore, in order to discover rules about candidate toxicophores, we generated fragments from the training set itself. We took each structure in the data set and removed the C-C and C-H bonds to generate a set of naïve fragments which were then matched against all the structures in the data set.

Taking the fragments occurring most frequently in the data set as parameters for generating models with FormRules resulted in simple, interpretable models in which SARs appear in the rules. The SARs are simple enough to be related directly to toxicophores or Derek alerts.

In addition, rules which cannot be linked to Derek alerts can be rationalised and represent generic knowledge not yet captured by Derek, e.g. the generally deactivating effect of carboxylic acid substituents.

In the model shown below, 7 out of 9 sub-models can easily be related to Derek alerts. The model was built from 6508 compounds using 150 fragments; the model took less than 2 minutes to build using a PC with 3GHz Pentium processor running Windows XP sp 3; FormRules SubModels are compared with alerts from the Derek Nexus 2.0 knowledge base.

CONCLUSIONS

FormRules’ neuro-fuzzy logic engine is able to produce models of toxicity data in which key SARs are discovered.

The most successful approach – producing the most interpretable models – is to take a data set and generate fragments therefrom, rather than to use off-the-shelf packages of fragment-based descriptors.

The models are not suitable as predictive models but can suggest SARs that can be further investigated and refined by human knowledge developers.