Self Organising Hypothesis Networks: Organising (Q)SAR knowledge

Thierry Hanser, Chris Barber, Edward Rosser, Jonathan Vessey, Samuel Webb, Stéphane Werner
Lhasa Limited, Granary Wharf House, 2 canal Wharf, Leeds, LS11 5PS

Background

In the context of regulatory risk management (Q)SAR models are expected to go beyond their role of classifier/estimator and provide wider assistance to human experts facing a difficult decision process. This requires not only good predictive performance but also a high degree of transparency. (Q)SAR transparency covers interpretability of the predictions (why did the model come to this conclusion?), prediction confidence level (is the model confident for this individual prediction?), supporting evidence (are there known examples of similar compounds for which we already know the behaviour?), and, understandable applicability domain (why is my compound in/out of the domain?).

Risk assessment

To address this challenge we have designed a new knowledge representation based on the concept of a hypothesis: a simple and interpretable knowledge unit. In this method, knowledge is broken down into a collection of hypotheses which are organised into a hierarchical network called Self Organising Hypothesis Network (SOHN). The resulting model is transparent and agnostic of its source of knowledge (learning process). When applied to building (Q)SAR models, this new approach provides accurate and yet transparent predictions along with individual prediction confidence levels and hence is well suited for risk assessment.

Methodology: Self Organising Hypothesis Network

The approach relies on 3 main steps:

1. Extract knowledge from one or more source of information (Learn)
2. Translate this knowledge into a unified representation independent of its origin (Unify)
3. Organise the knowledge according to its generalisation hierarchy (Organise)

For the final model to be transparent it must be composed of interpretable elements of knowledge. For that purpose we introduce the concept of a hypothesis. A hypothesis is a simple and interpretable knowledge unit that captures an elementary pattern of SAR behaviour. A hypothesis can be seen as a very local model; it defines a class of compounds that exhibit a SAR trend. In practice, hypotheses are either learned by a machine learning algorithm or elaborated by a human expert; they can take different forms depending on the information that they take into account (Figure 3).

Different hypotheses can be compared for their degree of generalisation (Figure 4, top), however comparing hypotheses of different types is more challenging (Figure 4, bottom). To solve this issue, the SOHN approach uses a reference dataset to define a generalisation order based on the coverage of the hypotheses within this dataset. The method is illustrated in Figure 5 with a reference dataset containing 10 examples (e1 to e10). If a hypothesis h1 covers examples e1 to e5 and a hypothesis h2 applies to the examples e1, e2 and e3 then it becomes intuitive to infer that h1 is more general than h2 since all the examples covered by h2 are also covered by h1.

Application: Sarah Nexus

To improve the safety in the context of mutagenicity risk assessment the international regulators introduced the ICH-M7 guidelines on genotoxic impurities, recommending the combined usage of an expert system and a (Q)SAR model. At Lhasa limited transparency in prediction has always been a priority and was our principle for developing Derek Nexus as an expert system for toxicity prediction. In response to ICH-M7, we developed Sarah Nexus a user friendly and fully featured (Q)SAR prediction application (Figures 13 and 14) to complement Derek Nexus. Sarah Nexus uses the SOHN methodology and has been trained for the Ames mutagenicity end-point; it is compliant with the OECD (Q)SAR principles[4] allowing for accurate and transparent predictions. Sarah Nexus has been favourably evaluated by the FDA[5]. Each prediction is supported by one or more self-explanatory hypotheses along with a confidence level and supporting evidence. This rich set of knowledge assists the toxicology expert in forming a well informed opinion and facilitates a safe decision process.

Conclusion

We have introduced a new paradigm in (Q)SAR that decouples the learning methodology from the final knowledge representation using a collection of hypotheses organised in a network. Using the most relevant part of this network as all holon models lead to accurate and highly interpretable predictions. Additionally this methodology provides access to a meaningful confidence metric at an individual prediction level which is critical in a regulatory context.

We have successfully integrated the SOHN methodology into a new prediction application Sarah Nexus. Ultimately the human expert remains the key asset in a risk assessment process. By providing transparent predictions Sarah and Derek Nexus combined offer valuable assistance in this decision process in line with the ICH-M7 and OECD guidelines.

References