Self-organizing High-dimensional QSAR Modeling from Noisy Data

Frank Lemke
KnowledgeMiner Software
Berlin, Germany
frank@knowledgeminer.com

Why QSAR Modeling?

Safety of pharmaceutical and chemical products with respect to human health and the environment has been a major concern for the public, regulatory bodies, and the industry, for a long time and this demand is increasing. Safety aspects start in the early design phases of drugs and chemical compounds and they end formally with the official authorization by national and international regulators. Traditionally, for decades, animal tests have been using as the preferred accepted tool - kind of Gold Standard, which, in fact, it is not - for testing harmful effects of chemicals on living species or the environment. Currently, in Europe only, about 10 million animals per year are (ab)used for laboratory experiments, and a lot of time and billions of Euros are spent into these experiments. So, we as consumers who use and value chemical products every day everywhere in some form are safe? No! Not really. About 90% of the chemicals on the market today have never been tested or have not been requested to be tested, officially. There is a simple reason, apparently: Despite the ethical issues of animal testing - it is estimated [1] that additional 10 - 50 million vertebrate animals would be required if all 150,000 registered substances would have to be tested in this traditional way - it is simply not possible to run animal tests for this amount of substances within reasonable time and cost constraints. Animal tests cannot do that. To solve this problem, there is a strong demand for alternative testing methods like QSAR [2] models to help minimizing and widely substituting animal tests in the future.

Many QSAR models for various chemical properties and biological endpoints have been published in the past 10 years, especially. However, most of them have been developed from a scientific viewpoint, only, and it is not clear if they are applicable for industrial and regulatory purposes. The current international research project ANTARES [3] funded by the European Commission targets this problem. It is searching and evaluating published QSAR models for a large number of endpoints using a set of quality, transparency, and reliability criteria important for identifying models, which can be used appropriately, and which are accepted by all parties involved, during official registration and authorization procedures of chemicals like the ongoing European initiative REACH [4, 5]. A new free online source about QSAR models for regulatory purposes developed by the "Mario Negri" institute, Politecnico di Milano, the US EPA, the UK Food and Environment Research Agency, and KnowledgeMiner Software will be available by the end of May 2011. This VEGA platform [6] will provide high-quality models for toxicity (carcinogenicity, mutagenicity, developmental toxicity, skin sensitization), eco-toxicity, and environmental endpoints in an innovative way.

The Modeling Approach

Predictive modeling of a biological activity from the molecular structure of chemical compounds can be seen as a complex, ill-defined modeling problem, which is characterized by a number of methodological problems:

- Inadequate a priori information about the system for adequately describing the inherent system relationships. Creating models for predicting harmful effects on human health and the environment is a highly interdisciplinary challenge. There is no domain knowledge available from any single domain that would solve the problem by theory.
- Possessing a large number of variables. A few hundred to a few thousand input variables are not uncommon in QSAR modeling.
- Noisy and few data samples in the range of tens to a few hundred data.
- Vague and fuzzy objects whose variables have to be described adequately. Experimental toxicity data are result of animal tests. Depending on the species used in an assay its inherent bio-variability can be quite high and can vary very much from species to species and from test to test. This translates into huge amount of noise in the experimental data used to build QSAR models.

A powerful modeling technology that addresses these problems by its design is *Self-organizing Networks of Active Neurons* based on the Group Method of Data Handling. Built on the principles of self-organization, it inductively develops, starting from the simplest possible model, optimal complex models that are composed of sets of self-selected relevant inputs (fig. 1). In this way, it performs both parameter and structure identification of a model and it solves the basic problem of experimental systems analysis of systematically avoiding overfitted models based on the data's information, only. Furthermore, the models are available analytically in form of linear or non-linear regression or difference equations. High-dimensional modeling from hundreds or thousands of input variables is another integrated part of Self-organizing Networks of Active Neurons that apply unique approaches to *multilevel self-organization and noise immunity* [7]. This leads to the concept of self-organizing high-dimensional modeling, which hides the complex processes of knowledge extraction, model development, dimension reduction, variables selection, noise filtering for avoiding overfitted models, and model validation from the user as a condition *for objectively developing reliable models from noisy data*.

Contest Results

For the SIAM SDM'11 QSAR Challenge we used our general-purpose predictive modeling and data mining tool KnowledgeMiner out of the box [8]. We also tested a new algorithm on cost-sensitive classification we are developing to see how it performs under real-world conditions. This algorithm also optimizes results of imbalanced class distributions as found in the challenge. The final solution submitted to the challenge is a model ensemble of two non-linear regression models obtained directly from the challenge data set of 837 samples and 242 descriptor variables. No prior dimension reduction, feature selection, data normalization or subdivision was used. All this is integrated in the knowledge extraction process of the tool.

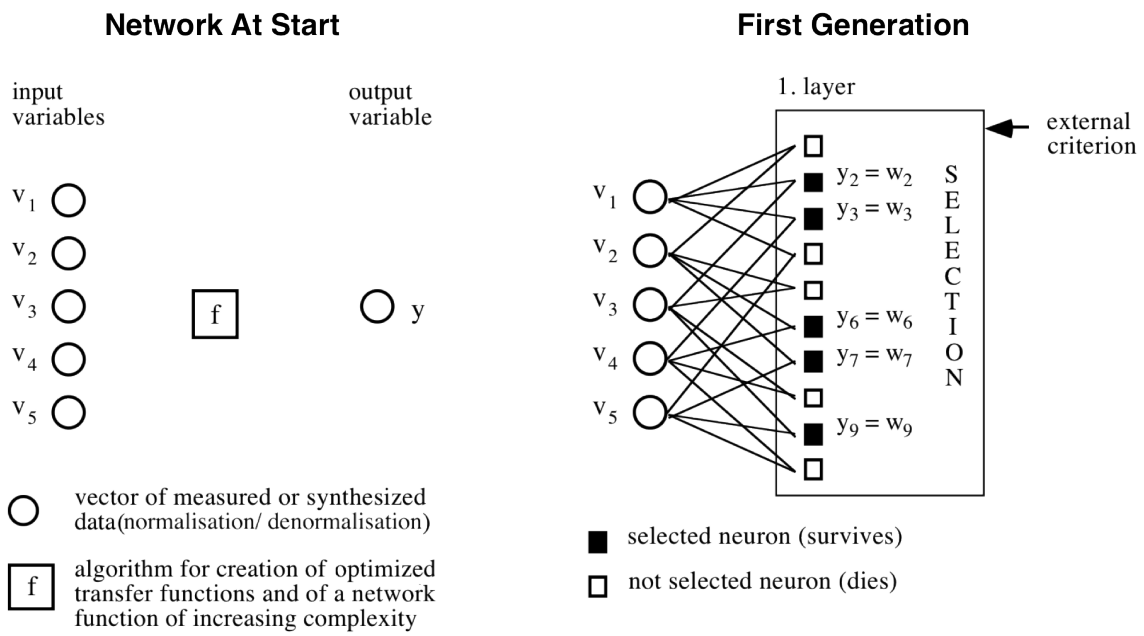
This may sound as a very time consuming modeling task, but since KnowledgeMiner is 64-bit parallel software, it actually is not. To self-organize a model from the entire challenge data set takes about 1 - 5 minutes on a 3 GHz 8-core Mac Pro running Mac OS X 10.6.

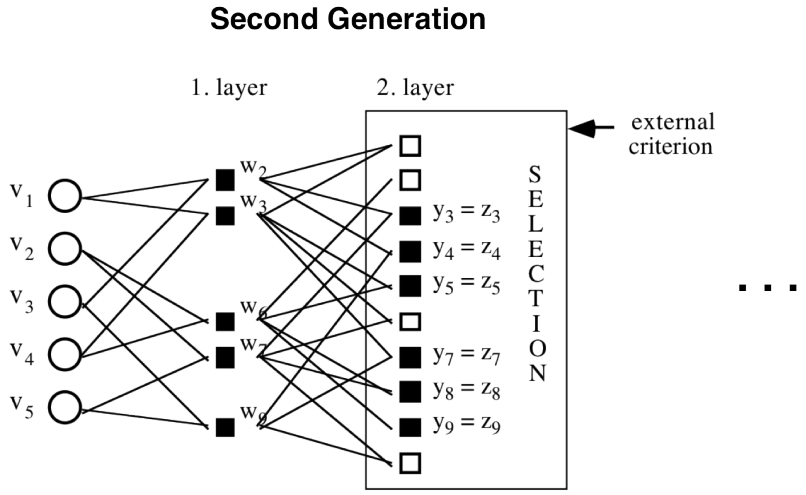
The first model is composed of 14 and the other of 15 self-selected relevant molecular descriptors, which join to a unique set of 19 descriptors:

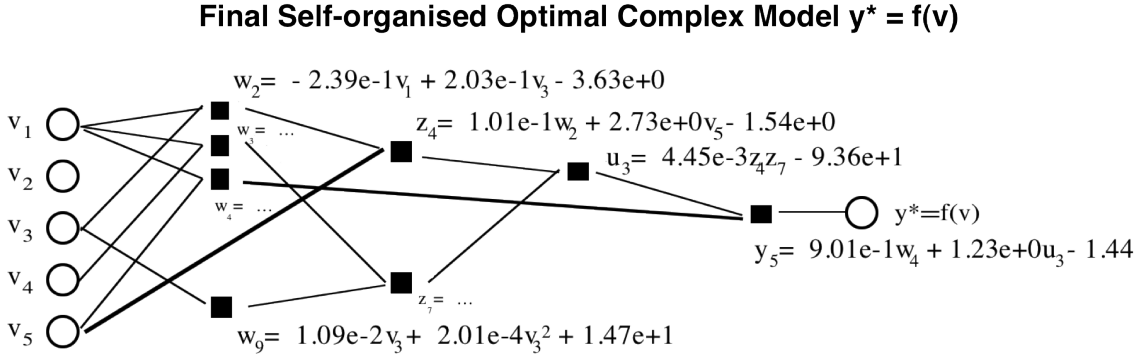
$$y_{Class} = f(x_i),$$

with i = {20, 29, 54, 70, 86, 98, 105, 138, 142, 144, 145, 147, 157, 159, 164, 174, 193, 204, 228}.

The sensitivity of the final combined model on the design data is 0.714, specificity is 0.707, the positive predictive value is 0.367, and the negative predictive value is 0.912. On the out-of-sample challenge test data set this model shows a sensitivity and specificity of 0.711 and 0.677, respectively.







- *Self-selected input variables*: v_1, v_3, v_4, v_5
- *Self-organised Transfer Functions*: Active Neurons
- *Self-organised Network Topology*: Analytical Optimal Complex Model

Figure 1. *Self-organization of a Network of Active Neurons.*

All models can be exported to Excel for further use.

A free sample model and a download of the software is available at [9] and [10] respectively.

References

[1] Hartung, T., Rovida, C.: Chemical regulators have overreached, *Nature* **460**, 1080-1081 (27 August 2009)

[2] Quantitative Structure-Activity Relationship Model, http://en.wikipedia.org/wiki/Quantitative_structure-activity_relationship

[3] Life+ Project ANTARES, http://www.antares-life.eu

[4] Registration, Evaluation, Authorisation and Restriction of Chemicals, http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm

[5] In silico methods for testing the toxicity of chemicals. An introduction. http://www.knowledgeminer.eu/pdf/introductory_leaflet.pdf

[6] Virtual Models For Evaluating The Properties Of Chemicals Within A Global Architechture. http://www.insilico.eu/

[7] Lemke, F.: Noise Immunity and Descriptive Power, http://www.knowledgeminer.eu/noise_immunity.html

[8] KnowledgeMiner, http://www.knowledgeminer.eu

[9] SIAM Challenge sample model, http://www.knowledgeminer.eu/bin/siam11_model1.zip

[10] KnowledgeMiner free version, http://www.knowledgeminer.eu/download.html