17th European Symposium on Computer Aided Process Engineering – ESCAPE17
V. Plesu and P.S. Agachi (Editors)

Neural network based predictions for the liquid crystal properties of organic compounds

Catalin Lisa[a], Silvia Curteanu[b]

[a,b] *"Gh. Asachi: Technical University of Iasi, Faculty of Chemical Engineering, Bd. D. Mangeron, No. 71A, IASI, 700050, ROMANIA, E-mail:* clisa@ch.tuiasi.ro, scurtean@ch.tuiasi.ro
For correspondence : scurtean@ch.tuiasi.ro or silvia_curteanu@yahoo.com

Abstract

This paper presents a new method of predicting the liquid crystalline behavior of some organic compounds, using feed-forward neural networks. The prediction of properties is correlated with molecular weight and a series of structural characteristics estimated by mechanical molecular simulation. An efficient genetic algorithm based method is used to determine optimal topology of the neural model.

Keywords: neural networks, genetic algorithms, liquid crystal properties.

1. Introduction

The design of materials possessing desired physical, chemical and biological properties is a challenging problem in the chemical, petrochemical and pharmaceutical industry. This involves modeling important interactions between basic structural units for property prediction as well as efficiently locating viable structures that can yield desired performance on synthesis [1].
The use of neural networks to the prediction of properties of organic compounds has as main advantage the fact that neural networks can simulate the nonlinear relationship between structural information and properties during the training process, and generalize the knowledge among homologous series without need

for theoretical formulas. The ability of neural networks is significant in determination quantitative structure-property relationship, because compounds with known properties can be used to train networks, so that, subsequently, properties of other compounds that can not be ascertained by experimentation can be determined [2].

2. Problem Statement, background

In our times, the reduction of the number of experimental trials represents a requirement that is more and more felt in the field of the study and analysis of chemical phenomena. Determination of the properties of some organic compounds based on their structures is a major research subject in computational chemistry. A common goal of materials science is the determination of relationships between the structure (microscopic, mesoscopic and macroscopic) of a material and its properties (mechanical, thermal, magnetic, optical, electrical, environmental and deteriorative). This information is crucial for engineering materials that provide a pre-determined set of properties [1]. The explosion in computational power of modern computers as well as their inexpensive availability has prompted the development of computer-assisted procedures for designing new materials to ease the protracted design, synthesis and evaluation cycle. Computational molecular design systems require the solution of two problems: the *forward* problem which predicts physical, chemical and biological properties from the molecular structure, while the *inverse* problem requires the identification of the appropriate molecular structure given the desired macroscopic properties.
The *property prediction methods* may be evaluated based on their classification as empirical, semi-empirical, theoretical and hybrid approaches. The empirical methods usually require extensive data collection and result in linear or simple nonlinear structure-property relations. Computations are very rapid at the expense of prediction accuracy. In addition, these methods require a specific functional form which may not always be available and the parameters determined by regression from the data. They are also computationally expensive, but provide excellent property estimations. Most approaches settle for the middle ground by utilizing simplified assumptions as those found in semi-empirical methods and hybrid approaches. These methods provide the best compromise between model development effort, computational time and property prediction accuracy. In this regard, neural network based methods offer advantages of ease of development and implementation, and execution speed, while maintaining a high degree of accuracy of predictions. Neural network based models are relatively model free, in the sense that the underlying functional form is not as rigorous as in the traditional model based methods. This adds to the generality of these methods.

Different machine learning algorithms, including hierarchical clustering, decision trees, k-nearest neighbours, support vector machines and bagging are used in structure prediction [3].

3. Paper approach

In the organic compounds' field, the efficient design of new materials requires the prediction of the compound properties and the selection of the best structure from all the potential possibilities. To solve this problem, a quantitative structure-property relationship is necessary and as a function of the investigated property some methods are given in the literature [4, 5]. One of the most interesting properties of organic compounds is the liquid crystalline (LC) behavior, because in this state the materials combine two essential properties of the matter: the order and the mobility. But, due to the complexity of the liquid crystalline phase, it is not at all easy to predict the occurrence of a mesophase. There are many methods of predicting the liquid crystalline behavior based on molecular, energetic or structure-property relationship models [6-8].
In this paper we used an organic compounds database [9] (122 in all) which includes a wide variety of azo aromatic compounds containing different units connected to the azo aromatic core. The present approach is an opportunity to prove the utility and the efficiency of the neural networks for classification problems, particularly for quantifying the relation structure – properties for some azo aromatic compounds. Simple neural networks and accessible methodologies provide good results in LC behavior predictions. A new genetic algorithm based method is used to design optimal topology for neural model. The prediction of properties is correlated with chemical structure, molecular weight and a series of structural characteristics estimated by mechanical molecular simulation.

3.1. Methodology

Feed forward neural networks represent a method for building models when a non-linear relationship is assumed [2]. The processing elements of a network (the neurons) are organized in layers and each neuron is linked to the neurons of the next layer. Typically, a feed-forward network consists of one input layer, some hidden layers and an output layer. In the training phase, the neural network learns the behavior of the process. The training data set contains both input patterns and the corresponding output patterns (also called target patterns). Neural training leads to finding values of connection weights that minimize differences between the network outputs and the target values. The most extensively adopted algorithm for the learning phase is the back-propagation algorithm. The purpose of developing a neural model is to devise a network (set of formulae) that captures the essential relationships in the data. These formulae are then applied to new sets of inputs to produce corresponding outputs. This is

called generalization and represents subsequent phase after training (validation phase). A network is said to generalize well when the input-output relationship found by the network is correct for input/output patterns of validation data that were never used in training the network (unseen data).

3.2. Experimental arrangement

The establishment of the numerical inputs for neural models (*molecular descriptors*) is a critical and difficult problem. This is due to the fact that the molecular descriptors must represent the molecular structural features related to the properties of interest as distinctly as possible. The prediction accuracy of neural networks depends heavily on the amount of correction between the molecular descriptors and the structural features. We used as molecular descriptors: length of the rigid core, length of the flexible core, total length, molecular diameter, molecular weight, ratio molecular diameter / total length. The molecular descriptors were estimated by mechanical molecular simulation using Hyperchem program. Concerning the liquid crystal behavior, we have coded with "1" the possibility to generate a mesophase and with "0" the crystalline or amorphous phases. This is the symbolic output of the model.

3.3. Case study

The combination of different structural units in a molecule gives rise to physical properties which are very important when designing new liquid crystals. For practical use, the materials should not only have the molecular structure suitable for inducing liquid crystal properties, but also an appropriate combination of physical properties for that application. The factors influencing the molecular unit are varied and include core units, connecting groups, terminal groups, lateral groups and lengths of flexible chains. All these structural factors affect the nature of interactions between liquid crystalline molecules and are very important for obtaining the adequate mesomorphic behavior. The organic compounds used in this paper have similar structures with small structural changes that allow a systematical analysis of the factors that influences liquid crystals properties and determination of some parameters that will be used in prediction with neural networks. Our database contains compounds with different units connected to the azo aromatic core such as CN, Br, variable length alkyl chains, ketones by means of ester or ether linking group.

3.4. Results & discussions

The feed-forward, multilayered neural network is the most used kind of neural networks because the simplicity of its theory, ease of programming and good results and because it is a universal function in the sense that if topology of the network is allowed to vary freely it can take the shape of any broken curve.

Firstly, the data are split into training and validation data sets because it is more important to evaluate the performance of the network on unseen data that training data. In this way, we can appreciate the most important feature of a neural model - the generalization capability.

One major problem in the development of neural network model is determining the network architecture, i.e. the number of hidden layers and the number of neurons in each hidden layer. We propose a genetic algorithm based method for detecting the optimal topology for a neural network that should approximate as well as possible the test data. The representation of solutions in chromosomes must simultaneously takes into account two problems: including the information on network topology (number of hidden layers, number of neurons in these layers) and including actually the connection weights and biases of the neurons, with the purpose of verifying the network training errors. All this information is coded by real numbers that is why we use the real encoding for the chromosome genes. The fitness function is equivalent in the present approach to calculating the mean square error for the test problem for the neural network represented by a certain chromosome. The chosen representation has both advantages and disadvantages. The advantage is the simplicity of the approach, as the genetic algorithm also accomplishes the finding of the optimum topology and the training of the neural network (determining the connection weights that allow approximating the test data). The disadvantage is represented by a long training time because of the big number of chromosome genes (information regarding the topology and the connection weights and the biases of the neurons). Details about our method is given in [10]. A MLP(4:42:14:1) is obtained, with MSE (Mean Squared Error) = 0.01831, E_p (percent error) = 1.1133 % and r (correlation) = 0.9885.

Table 1. Validation of the neural model, MLP(4:42:14:1)

Length of the rigid core	Length of the flexible core	Molecular diameter / total length	Molecular weight	LC	LC net
9.21	25.5	0.08	463	0	0
9.22	20.98	0.09	439	0	0
9.22	6.22	0.19	270	1	0
9.22	8.77	0.16	298	1	1
9.23	8.9	0.16	296	1	1
9.23	16.62	0.11	381	0	0
9.21	6.39	0.18	266	0	0
9.21	9.94	0.15	310	1	1
9.21	20.61	0.10	439	1	1
9.21	11.69	0.14	360	1	0
9.21	17.24	0.11	431	0	0
9.21	15.2	0.12	404	0	0

The predictions of the neural network MLP(4:42:14:1) on the training data were compared to the experimental ones in order to verify how the network projected

has learned the behavior of the process. The correlation between the two sets of data, 0.99, and the probability of a correct answer of 99 % show a good concordance between the model and the experimental results. A key issue in neural network based process modeling is the robustness or generalization capability of the developed models, *i.e.* how well the model performs on unseen data. Thus, a serious examination of the accuracy of the neural network results requires the comparison with experimental data, which were not used in the training phase (previously unseen data). The predictions of the networks on validation data are given in Table 1 (LC net compared to LC obtained experimentally). Cells marked in black represent wrong predictions of the network. In the validation stage, the probability of a correct answer of MLP(4:42:14:1) was 83.33 %, that is a good performance of the designed network. Consequently, a feed-forward network MLP(4:42:14:1) can predict satisfactory the LC behavior of the compounds.

4. Conclusions/Remarks/future work

The prediction of the mesophase occurrence with machine learning methods as well as the choice and the codification (numerical and nominal) of different sets of parameters which characterize the structure and the behaviour of the azo aromatic compounds represent a new approach in the field. Neural network based method proved to be able to appreciate the liquid crystalline behaviour with small errors, so it represents an effective tool for structure – properties prediction. Simple feed-forward neural network with optimal topology developed within a genetic algorithm based procedure was used in this paper. We intend in our future research to extend the database including other types of organic compound and to use different machine learning methods such categorization algorithms.

References

1. V. Venkatasubramanian, K. Chan, and J.M. Caruthers, Computers Chem. Engng,. 18 (1994) 833.
2. J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, 2nd Edition, Wiley-VCH, Weinheim, 1999.
3. T. Kleinoder, S. Spycher, A. Yan, Prediction of properties of compounds, in Chemoinformatics – A Textbook, Wiley-VCH, Weinheim, 2003.
4. N.K. Roy, W.D. Potter, D.P. Landau, IEEE Transactions on Neural Networks, 17 (2006) 1001.
5. N.K. Roy, W.D. Potter, D.P. Landau, Appl. Intell., 20 (2004) 215.
6. C. Yan, V.Honavar, D. Dobbs, Neural Comput. & Applic., 13 (2004) 123 129.
7. G.A.Landrum, H.Genin, J. Solid State Chem., 176 (2003) 587.
8. P.Villars, Eng. Appl. Artif. Intell., 13 (2000) 497.
9. ***LiqCryst Online, Liquid Crystal Group Hamburg.
10. S. Curteanu, F. Leon, Int. J. Quant. Chem., (2006) in press.