Macromolecular Research, Vol. 10, No. 1, pp 13-17 (2002)

Influences of the Input on ANN and QSPR of Homopolymers

Hong Sun, Yingwu Tang*, and Guoshi Wu

Department of Chemistry, Tsinghua University, Beijing, P.R.China, 100084

Received Aug. 13, 2001 ; Revised Jan. 29, 2002

Abstract: An artificial neural network (ANN) was used to study the relationship between the glass transition temperature (T_g) and the structure of homopolymers. The input is very important for the ANN. In this paper, six kinds of input vectors were designed for the ANN. Of the six approaches, the best one gave the T_g of 251 polymers with a standard deviation of 8 K and a maximum error of 29 K. The trained ANN also predicted the T_g of 20 polymers which are not included in the 251 polymers with a standard deviation of 7 K and a maximum error of 21 K.

Keywords : artificial neural network, QSPR, input vector, glass transition temperature.

Introduction

Several methods have been developed to study the quantitative structure-property relationship (QSPR) of polymers, such as the group contribution approach (GCA),[1] the molecular connectivity indices approach (MCIA),[2] the artificial neural network approach (ANNA),[3-5] etc.

An artificial neural network (ANN) is a machine designed to model the way in which the brain performs a particular task or function of interest. Because of its capability for parallel computing and the information storage, the ANN can be used for nonlinear vector mapping, pattern recognition, classification and so on.[6] The ability of ANN to perform nonlinear vector mapping was used to study the QSPR of polymers.[3-5]

Since the QSPR of polymers is very complicated and the structural parameters that affect the properties are not independent but interact, the structure - property relationship is not linear. Therefore the ANN was used to find the nonlinear relationship between the glass transition temperature (T_g) and the structure of homopolymers in this paper. Six kinds of input vectors are designed for the ANN. The results from these six inputs methods were used to investigate the influences of the input method on the ANN and the QSPR of homopolymers.

Methods

"NN-T_g". A Multi-Layer Feed-Forward Neural Network (MLFN) used to analyze the QSPR is referenced to as "NN-T_g" in this paper. "NN-T_g" was constructed using the neural network software packages in Matlab. Figure 1 illustrates the architecture of "NN-T_g", which included an input layer, a hidden layer and an output layer.

Neurons in the input layer act only as buffers for distributing the input signals x_i ($i = 1 \sim R$) to neurons in the hidden layer. Each neuron j ($j = 1 \sim S$) in the hidden layer sums up its input signals x_i after weighting them with the strengths of the respective connections ω_{ij} ($i = 1 \sim R$, $j = 1 \sim S$) from the input layer and computes its output a_j ($j = 1 \sim S$) as an activation function f of the sum, viz.

$$a_{ij} = f\left(\sum_{i=1}^{R} \omega_{ij} x_i\right), \quad j = 1, 2, \ldots, S \tag{1}$$

where f is a hyperbolic tangent function

$$f(x) = \frac{2}{1 + exp(-2x)} - 1 \tag{2}$$

The output y of neurons in the output layer is computed in a similar way, but with the activation function f as a linear function $f(x) = kx$.

The backpropagation (BP) algorithm, a gradient descent algorithm, was chosen as the training algorithm for "NN-T_g". The number neurons, S, in the hidden layer was selected during the training process.

Input Vector x. The T_g for 271 homopolymers (Table I) were studied.[7] 20 polymers out of 271 polymers were used for the testing set (Table II) and the rest as the training set. The polymers' T_g ranged from 183 K to 593 K.

The repeat unit of these polymers is (-(CH_2-CHR)$_n$-) with different substituents R, so the group concept in GCA[1] was used to describe the polymer structure. For example, the poly

*e-mail : ywtang@mail.tsinghua.edu.cn

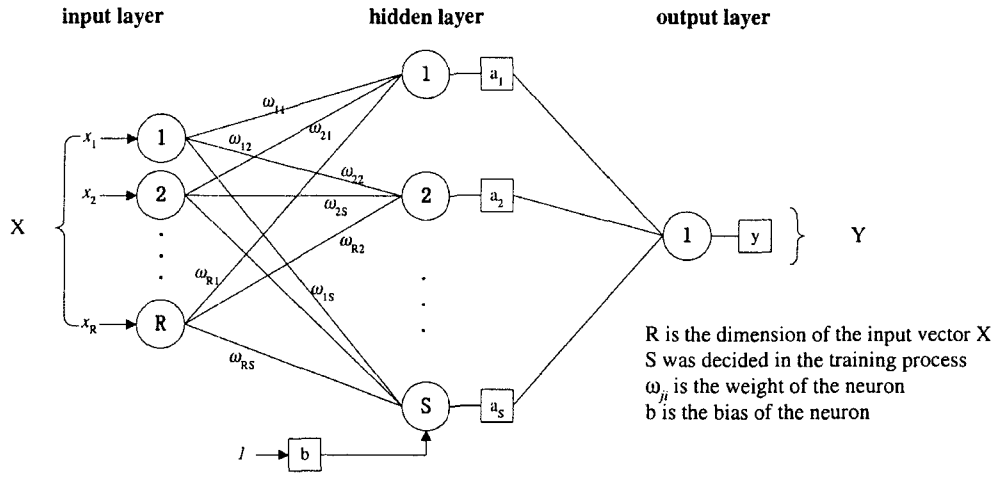

Figure 1. "NN-T_g" architecture.

Table I. Classes of 271 Polymers

Poly(acrylate) (65)	Poly(vinyl esters) (30)	Polyolefins (33)
Poly(vinyl ethers) (15)	Poly(styrene) (106)	Poly(vinyl ketones) (10)
Poly(acrylamide) (10)	Others (2)	

*The value in parenthesis represents the number of polymers in that class.

Table II. Predicted and Experimental T_g of 20 Polymers in the Testing Set

Polymers	T_g (exp.)	T_g' (predicted)	T_g'-T_g
Poly(4-cyanophenyl acrylate)	363	369	6
Poly(3-nitrobenzoyloxy ethylene)	366	364	-2
Poly(4-phenylbenzoyloxy ethylene)	358	363	5
Poly(3-o-methylphenyl propylene)	353	362	9
Poly(vinyl chloride)	354	351	-3
Poly(4-methoxy-2-methyl styrene)	358	356	-2
Poly(2,5-dichloro styrene)	379	374	-5
Poly(3-methyl styrene)	370	374	-4
Poly(2-methoxymethyl styrene)	362	367	5
Poly(4-methoxy styrene)	362	372	10
Poly(4-propionyl styrene)	375	371	-6
Poly(4-propoxycarbonyl styrene)	365	366	1
Poly(2-benzoyloxymethyl styrene)	345	351	6
Poly(5-bromo-2-methoxy styrene)	359	361	2
Poly(4-butyryl styrene)	347	368	21
Poly(4-ethoxycarbonyl styrene)	367	367	0
Poly(4-ethoxy styrene)	359	362	3
Poly(4-isobutoxycarbonyl styrene)	363	367	4
Poly(2-pentyloxycarbonyl styrene)	365	366	1
Poly(tert-butoxy ethylene)	361	359	-2

(4-butoxycarbonylphenyl acrylate) (R=COOPhCOOC$_4$H$_9$) can be described by four sequential groups -COO-, -Ph-, -COO- and -C$_4$H$_9$. Because the longest substituent in the 271 polymers included only four groups, R is divided into four

groups R_1, R_2, R_3 and R_4, where R_2, R_3 and R_4 may equal zero. The number m_i ($i = 1 \sim 4$) of different groups in each R_i ($i = 1 \sim 4$) was 18, 32, 24 and 10, so the sum was 74. Each of these groups were identified by r_{ji} ($j = 1 \sim m_i$, $i = 1 \sim 4$). If 0 and 1 are used to represent not including or including the group, the input dimension is 74. The finite data is not good for the ANN training so several methods were developed to describe r_{ji}.

The first approach is called the "sorting order method". First, the average T_g for the polymers who include the group r_{ji} was calculated. Then the average were sorted and given a value r_{ji}^1:

$$r_{ji}^1 = -1 + \frac{2}{m_i} \times P_{r_{ji}} \tag{3}$$

where $P_{r_{ji}}$ is the order number of r_{ji} in R_i. The structure of each polymer can be expressed by the r_{ji}^1 which gives a four-dimensional input vector $V_1 = \{r_{ji}^1(k), k = 1,2,3,4\}$. V_1 was then used as the input to train the network.

The second approach is called the "averaging method". The average r_{ji}^2 ($j = 1 \sim m_i$, $i = 1 \sim 4$) of the T_g for the polymers who include the group r_{ji} was calculated using:

$$r_{ji}^1 = \sum_{k=1}^{N_{r_{ji}}} r_{jik}^0 / N_{r_{ji}} \tag{4}$$

where $N_{r_{ji}}$ is the number of polymers who include the group r_{ji} and r_{jik}^0 are the T_g of the polymers who include r_{ji}. The structure of each polymer can be expressed by r_{ji}^2 which gives a four-dimensional input vector V_2. Sometimes, different groups have very similar averages, the second center moment (SCM) and the third center moment (TCM) were used.[8]

$$r_{ji}^3 = \sum_{k=1}^{N_{r_{ji}}} (r_{jik}^0 - r_{ji}^2)^2 / N_{r_{ji}}, j = 1 \sim m_i, i = 1 \sim 4 \tag{5}$$

$$r_{ji}^4 = \sum_{k=1}^{N_{r_{ji}}} (r_{jik}^0 - r_{ji}^2)^3 / N_{r_{ji}}, j = 1 \sim m_i, i = 1 \sim 4 \tag{6}$$

The combination of r_{ji}^2 and r_{ji}^3 gives an eight-dimensional input vector V_3. The combination of r_{ji}^2, r_{ji}^3 and r_{ji}^4 gives a twelve-dimensional input vector V_4. V_2, V_3 and V_4 were then used as inputs to train the network.

The third approach is called the "connecting method". A three-dimensional data set r_{jk}^5 ($j = 1 \sim m_i$, $i = 1 \sim 4$, $k = 1 \sim 3$) was introduced to represent the connections between R_1 and R_2, R_2 and R_3, and R_3 and R_4 for $k = 1$ to 3. As with equation (4), r_{jk}^5 is also expressed by the average T_g for polymers who include the connection. The combination of r_{ji}^2, r_{ji}^3, r_{ji}^4 and r_{jk}^5 gives a fifteen-dimensional input vector V_5. The combination of r_{ji}^1 and r_{jk}^5 gives a seven-dimensional input vector V_6. V_5 and V_6 were also used as inputs to train the network.

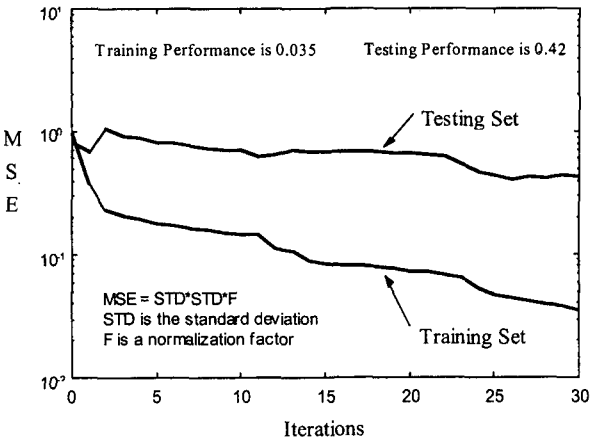

Figure 2. Training and testing of "NN-T_g".

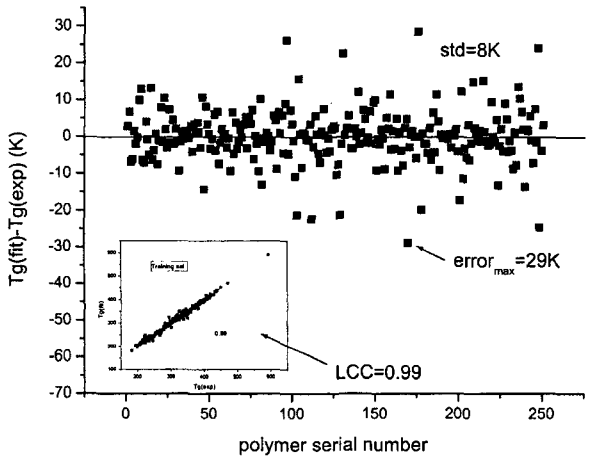

Figure 3. Training set errors with V_5 as the input for "NN-T_g".

Results and Discussion

Results of "NN-T_g" with V_1, V_2, V_3, V_4, V_5 and V_6 as Inputs. Each one of V_1, V_2, V_3, V_4, V_5 and V_6 were used separately as an input for "NN-T_g" and T_g as an output. The training set was used to train the network with the testing set used to test it. The performance is shown in Figure 2 where STD is the standard deviation calculated using:

$$STD = \sqrt{\sum_{i=1}^{N} (T_{gi}(fit) - T_{gi}(exp))^2 / N} \tag{7}$$

where N is the number of polymers in the training set or testing set. For the ANN, because the T_g are larger than 1, the T_g were normalized with a normalization factor F. The performance in Figure 2 is equal to the product of F and the square of STD. With increasing number of iterations, the errors of the training set and the testing set decrease and finally reach the minimum. Figure 3 (or Figure 4) compares the fitted (or predicted) T_g with the experimental data for the

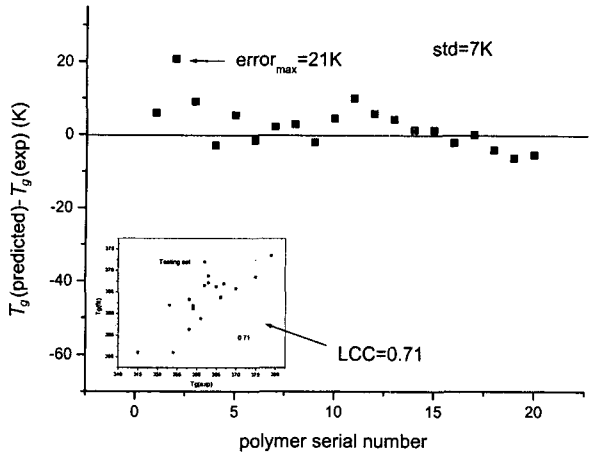

Figure 4. Testing set errors with V_5 as the input for "NN-T_g".

training set (or the testing set) when V_5 is the input of "NN-T_g". The trained ANN not only quantifies the relationship between the T_g and the polymer structure, but also predicts the T_g of the polymers not included in the training set. The testing set was used to confirm that the trained ANN could be used for prediction. The testing set results are not so good as the training set results, but the precision shows that the trained ANN can be used to predict T_g. Therefore the network can give the right function to map the relationship of the input to the output which can be used to predict T_g of polymers which are not included in the training set.

The results using V_1, V_2, V_3, V_4, V_5 and V_6 as inputs to train "NN-T_g" are listed in Table 3. STD is the standard deviation and ME is the maximum error for the polymers. LCC is the linear correlation coefficient for the results in Figure 3 or Figure 4.

Comparison of V_1, V_3 and V_4 with V_2. All of V_1, V_3 and V_4 are derived from V_2, therefore, their original information is same but the different mathematical tools retrieve different information. V_1 adds the sorting order of the average, V_3 adds SCM and V_4 adds SCM and TCM. The main purpose is to prevent giving a same value for different groups.

Table III gives a concrete example. The three groups 2,4-phenyl (G1), 4-phenyl (G2) and iso-propyl (G3) all have r_{ji}^1 equal to -0.25 as in V_1. All of the three groups have the same description, thus the ANN can not distinguish them. In V_3, the introduction of SCM distinguishes G3 from G1 and G2 since r_{ji}^3 is different. In V_4, the introduction of SCM and TCM separates the three groups by the values of r_{ji}^3 and r_{ji}^4.

Compared to V_3 and V_1, V_4 gives more information about the different groups. In Table IV, the results using V_1 are better than those using V_2, the results using V_3 and V_4 are better than those using V_2, and the results using V_4 are better than those using V_3. These indicate that the more information about the differences between the groups improve the results.

Comparison of V_5 and V_6 with V_1, V_2, V_3 and V_4. The

Table III. Group Data

Groups	r_{ji}^1	r_{ji}^3	r_{ji}^4
2,4-phenyl	-0.25	0.3	0
4-phenyl	-0.25	0.33	0.28
iso-propyl	-0.25	0.17	-0.12

Table VI. Training Set and Testing Set results for "NN-T_g"

Input	Training Set			Testing Set		
	STD*	ME*	LCC*	STD	ME	LCC
V_1	15	59	0.94	28	56	0.68
V_2	26	111	0.83	15	27	0.32
V_3	18	55	0.91	12	24	0.40
V_4	11	41	0.96	10	18	0.63
V_5	8	29	0.99	7	21	0.71
V_6	5	21	0.99	27	43	0.71

*STD = standard deviation.
*ME = maximum error.
*LCC = linear correlation coefficient.

information about the atoms connecting two groups is added in V_5 and V_6, so V_5 and V_6 differ from V_1, V_3 and V_4 because this information is new and cannot be derived from V_2. With this connection information, the STD and ME of V_5 are less than that of V_4 (and that of V_6 are less than that of V_1) so the LCC of the training set is improved from 0.96 to 0.99 (from 0.63 to 0.71 for the testing set). Therefore the connection information is very useful and interactions exist between the groups.

When V_5 and V_6 are compared, the results with the training set are similar, but the results of V_6 with the testing set are much better than with V_5. This indicates that less information is introduced by the sorting order than by SCM and MCE. Comparing V_3 and V_4 to V_1 also shows that SCM and TCM give more useful information than the sorting order.

Six Input Vectors. The problem can be expressed by the function $Y = F(X)$, where X represents the structural parameters and Y represents the property (T_g). Supposing $X = (x_1, x_2, \cdots, x_n)$, the information of the groups is expressed by X_1 ($X_1 = (x_1, x_2, \cdots, x_m) \subset X$), the information of the interactions is expressed by X_2 ($X_2 = (x_{m+1}, x_{m+2}, \cdots, x_q) \subset X$), and the information that cannot currently be determined is expressed by X_3 ($X_3 = (x_{q+1}, x_{q+2}, \cdots, x_n) \subset X$). Assuming that and $X = X_1 \cap X_2 \cap X_3$ and X_1, X_2 and X_3 are independent then

$$\begin{aligned} Y = F(X) &= F(X_1 \cap X_2 \cap X_3) \\ &= F_1(X_1) + F_2(X_2) + F_3(X_3) \end{aligned} \tag{8}$$

If $F_2(X_2)$ and $F_3(X_3)$ can be neglected, then $Y' = F_1(X_1)$. The methods using input vectors V_1, V_2, V_3 and V_4 just replace Y with Y'. Different mathematical expression for X_1

will produce different results. Therefore, more information identifying the differences between groups will produce improved results.

If just $F_3(X_3)$ is neglected, then $Y'' = F_1(X_1) + F_2(X_2)$. The results using input vectors V_5 and V_6 just replace Y with Y''. The errors in Y'' are much smaller than in Y', so $F(X_2)$ is an important part of Y and cannot be ignored.

Conclusion

The ANN was used to study the relationship between T_g and the structure of homopolymers. The results indicate that the polymer structures can be described using the groups and the connection information of the groups. Six kinds of input vectors were designed for the ANN based on different ways of transferring the qualitative structural description into a quantitative description. The results demonstrate that ANN input must be carefully designed to obtain the maximum amount of information from the limited data to model the polymer structure.

References

(1) van D.W. Krevelen, *Properties of Polymers*, Third edition, Amsterdam, Elsevier, 1990.
(2) J. Bicerano, *Prediction of Polymer Properties*, Marcel Dekker, New York, 1993.
(3) C. W. Ulmer II, D. A. Smith, and B. G. Sumpter, *et al.*, *Comput. Theor. Polym. Sci.*, **8**(3/4), 311 (1998).
(4) B. G. Sumpter and D. W Noid, *Macromol. Theory Simul.*, **3**, 363 (1994).
(5) N. K. Ebube, G. O. Ababio, and C. M. Adeyeye, *International Journal of Pharmaceutics*, **196**, 27 (2000).
(6) S. Haykin, *Neural Networks - a Comprehensive Foundation*, Macmillan College Publishing Company, New York, 1994.
(7) J. Brandrup, *Polymer Handbook*, Third edition, New York, 1989.
(8) B. R. Bhat, *Modern Probability Theory - an Introductory Textbook*, Second edition, New York, 1985.