DOI 10.1007/s00894-010-0885-3

ORIGINAL PAPER

Theoretical study on modeling and prediction of optical rotation for biodegradable polymers containing α-amino acids using QSAR approaches

Shadpour Mallakpour · Mehdi Hatami · Hassan Golmohammadi

Received: 20 July 2010 / Accepted: 20 October 2010 / Published online: 9 November 2010

Abstract The main purpose of the present study was modeling and prediction of the optical rotation ($[M]_D$) of some biodegradable polymers containing α-amino acids using quantitative structure-activity relationship (QSAR) approaches. In order to attain this goal, the optical rotation of a collection of 53 polymers was selected as a data set. The data set was randomly divided into three sections, training, test and external validation sets. By using dragon software, various descriptors were calculated for all molecules in the data set. The important descriptors were selected applying genetic algorithm-partial least squares (GA-PLS) method. Then an artificial neural network (ANN) was written with MATLAB 7 and used these descriptors as inputs and its output was optical rotation of desired polymers. Then, the constructed network was used for the prediction of ($[M]_D$ values of validation set. The squared correlation coefficient R^2 values of the ANN model for the training, test and validation sets were 0.998, 0.996 and 0.996 respectively. The results showed the ability of developed ANN to predict optical rotation of various polymers.

S. Mallakpour (✉) · M. Hatami
Organic Polymer Chemistry Research Laboratory,
Department of Chemistry, Isfahan University of Technology,
Isfahan 84156-83111, Islamic Republic of Iran
e-mail: mallak@cc.iut.ac.ir

S. Mallakpour
e-mail: mallakpour84@alumni.ufl.edu

S. Mallakpour
Nanotechnology and Advanced Materials Institute,
Isfahan University of Technology,
Isfahan 84156-83111, Islamic Republic of Iran

H. Golmohammadi
Department of Chemistry, University of Mazandaran,
47415 Babolsar, Islamic Republic of Iran

Keywords Artificial neural network · Genetic algorithm · Optical rotation · Partial least squares · Quantitative structure-activity relationship

Introduction

Optical rotation, the rotation of plane-polarized light by chiral species, takes place because such samples demonstrate differing refractive indices for left and right circularly polarized light [1, 2]. This phenomenon is referred to as circular birefringence and is reliant on the propagation of plane polarized light through a chiral medium. The optical rotation of a chiral molecule depends on its absolute configuration. In principle, absolute configurations of chiral molecules should be derivable from their optical rotations. In practice, optical rotations are seldom used for this purpose. This is attributable to the deficiency of practical, unfailing algorithms concerning optical rotation and absolute configuration.

The molecular optical rotation of a substance $[M]_D$ can be calculated as expressed in Eq. 1:

$$[M]_D = \sum_i f_i \sum_k K_{XY} \sin \alpha_k \tag{1}$$

where f_i is the population of the conformer i, α_k is the dihedral angle formed by the four consecutive atoms (Fig. 1) for each conformer i and K_{XY} is the constant of rotation for each type of dihedral angle. The sums are extended to all the dihedral angles (k) present in each conformer and to all of the conformers (i).

Understanding and predicting the molecular properties of chiral molecules is a principal aim of organic chemistry. One of the focuses has been on the synthesis of these types of molecules, among the ability to predict and control their

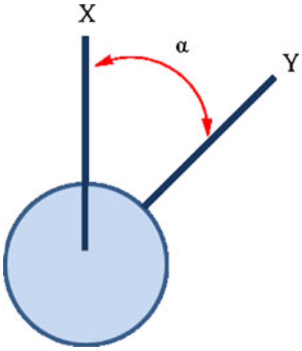

Fig. 1 X-C-C-Y, dihedral angle contribution to molecular optical rotation

properties. Chiral molecules are chiefly found in pharmaceutical chemistry, where it is sometimes necessary to control the absolute configuration of the molecule. Experimentally, reliable determination of the absolute configuration of a chiral molecule (optical rotation) is very expensive and time consuming, and is not guaranteed to be successful. Therefore, the development of a theoretical method such as quantitative structure–property/activity relationship (QSPR/QSAR) appears to be valuable to estimate the optical rotation values. The advantage of this approach over other methods lies in the fact that the descriptors used can be calculated from structure alone and are not dependent on any experiment properties. Once the structure of a compound is known, any descriptor can be calculated. Thus, once a reliable model is established, we can use this model to predict the property of a compound, whether it was obtained or not.

A number of theoretical attempts to predict the optical rotation of compounds have been performed. Nunez Miguel et al. [3] predicted optical rotation for a series of cyclohexane derivatives using molecular mechanics methods MM2 and MM3. Ruud and coworker [4] studied optical rotation on difficult systems by density-functional and coupled-cluster methods. Stephens et al. [5] developed a method using ab initio density functional theory (DFT) to estimate optical rotation of six 6,8-dioxabicyclo[3.2.1] octanes. The last few years have seen an increasing interest in the theoretical calculations [6–9] especially in optical rotation [10–14].

Recently, artificial neural networks (ANNs) have been used to a wide variety of chemical problems such as spectral analysis [15], prediction of dielectric constant [16] and mass spectral search [17]. ANNs have been applied to QSPR/QSAR analysis since the late 1980s due to its flexibility in modeling of nonlinear problems, mainly in response to increase accuracy demands; they have been widely used to predict many physicochemical properties [18–22]. In this work QSAR studies were carried out for the first time to find out the correlation between structural features of biodegradable polymers with their physicochemical properties such as optical rotation.

Modeling methodology

Data set

The data set of optical rotation was taken from the values reported by Mallakpour and coworkers [23–32]. The name of molecules in the data set including biodegradable polymers is shown in Table 1. The optical rotations of all macromolecules included in data set were obtained under the same conditions by a Jasco Polarimeter. The data set was randomly divided into three groups including training, test and validation set, which consists of 37, 8 and 8 molecules, respectively. The training and test sets were used to build and optimize the QSAR model and the external validation set was used to evaluate the prediction power of the obtained model.

Descriptor calculation

The molecular descriptors were encoded numerically with molecular features of the interested molecule. The built model performance and the accuracy of the results are robustly maintained by the way in which the structural representation was performed. It is impossible to calculate descriptors directly for an entire molecule because all the polymers have wide distribution of molecular weight and possess high molecular weight. As we know, if the molecular weight is high enough, the terminal groups hold only a very small proportion in a polymer and its effect on the properties can be ignored. Molecular descriptors calculated directly from the structure of the repeating units can be used for the study of QSARs for polymers, since all the properties depends on the chemical structure of the polymer molecule, and all these structures were conditioned by the repeating unit structures. Therefore, we adopted this method and concentrated on the following model to calculate molecular descriptors. The structures for polymers were endcapped with the last group of the opposing side. In the next step, the molecular structure of monomers used in the polymerization process, were used to determine the molecular descriptors of polymers. After providing the data set, all monomers were drawn into Hyperchem software and then pre-optimized using AM1 molecular mechanics force field [33]. A more precise optimization is then done with the semiempirical PM6 method in Mopac (2009) [34]. Since the calculated values of the electronic features of the molecules will be influenced by the conformation used, in the current research we made an attempt to use the most stable conformations. To avoid the local stable conformations of the compounds, geometry optimization was run many times with different starting points for each molecule, and the conformation with the lowest energy was considered for the calculation of the electronic properties. In a

Table 1 Data set and corresponding observed and predicted values of optical rotation of polymers

Number	Name of monomers of polymers	$[M]_D$ (EXP)	$[M]_D$ (PLS)	$[M]_D$ (ANN)
	Training set			
1	a and bisphenol A	-31.00	-33.48	-31.31
2	a and 4,4'-dihydroxydiphenyl sulphide	-40.90	-42.87	-41.41
3	a and 1,4-dihydroxybenzene	-9.10	-8.29	-8.69
4	a and bisphenyl-2,2'-diol	-9.60	-10.42	-9.63
5	b and phenol phthalein	-27.80	-29.64	-26.77
6	b and bisphenol-A	-17.00	-14.74	-17.76
7	b and 4,4'-hydroquinone	-21.20	-22.06	-20.34
8	b and 1,8-dihydroxyanthraquinone	-13.40	-16.19	-14.15
9	b and dihydroxy biphenyl	-20.60	-24.82	-20.02
10	c and 1,6-hexamethylenediamine	-29.30	-23.56	-28.26
11	c and 4,4'-sulfonyldianiline	-26.20	-31.58	-25.47
12	d and 4,4'-sulphonyldianiline	-14.20	-11.12	-15.49
13	d and 4,4'-Diaminodiphenylmethane	-21.50	-16.33	-20.13
14	d and 1,4-phenylenediamine	-23.20	-18.08	-24.27
15	d and 4,4'-diaminobiphenyl	-35.30	-38.93	-34.05
16	e and phenolphthalein	-9.20	-11.79	-8.96
17	e and 1,4-dihydroxybenzene	-3.25	-5.77	-3.70
18	e and 4,6-dihydroxypyrimidine	-3.25	-3.57	-3.09
19	e and 2,4'-dihydroxyacetophenone	-4.38	-7.00	-4.92
20	f and 4,4'-sulphonyldianiline	-55.50	-50.18	-57.04
21	f and 4,4'-diaminodiphenyl methane	-48.10	-43.16	-49.36
22	f and 4,4'-diaminodiphenylether	-19.20	-24.45	-19.66
23	f and *m*-phenylenediamine	-18.30	-13.01	-17.63
24	f and 4,4'-diaminobiphenyl	-40.20	-41.12	-41.32
25	g and phenolphthalein	-7.20	-12.72	-7.85
26	g 1,4-Dihydroxybenzene	-3.25	-5.07	-3.77
27	g and 4,6-dihydroxypyrimidine	-3.27	-5.84	-3.48
28	g and 2,6 dihydroxytoluene	-5.56	-3.13	-5.86
29	h and 4,4'-sulphonyldianiline	-86.10	-63.34	-83.63
30	h and 1,3-phenylenediamine	-50.90	-56.89	-52.13
31	i and 4,4'-diaminodiphenyl methane	-53.20	-42.30	-51.80
32	i and 4,4'-diaminodiphenylether	-16.24	-10.52	-15.17
33	i and *p*-phenylenediamine	-9.76	-15.91	-9.45
34	i and *m*-phenylenediamine	-9.26	-15.82	-9.15
35	and 2,4-diaminotoluene	-1.00	-3.27	-1.08
36	j and *p*-phenylenedi-amine	-13.00	-17.54	-13.77
37	j and 4,4'-diaminodiphenylether	-12.60	-17.65	-13.36
	Test set			
38	a and 1,4-dihydroxybenzene	-26.70	-23.73	-28.59
39	b and 1,4-dihydroxyanthraquinone	-18.40	-15.82	-17.21
40	c and benzidine	-8.60	-10.63	-8.01
41	d and 4,4'-Diaminodiphenylether	-29.20	-35.77	-28.21
42	e and 4,4'-dihydroxydiphenyl sulfide	-8.26	-7.24	-8.89
43	g and 4,4'-dihydroxydiphenyl sulphide	-7.26	-5.71	-7.63
44	h and 4,4'-Diaminodiphenylmethane	-60.30	-45.79	-62.47
45	i and 4,4'-diaminobiphenyl	-12.12	-18.10	-12.85
	Validation set			
46	a and 2,6 dihydroxy toluene	-13.20	-16.01	-12.28

Table 1 (continued)

Number	Name of monomers of polymers	$[M]_D$ (EXP)	$[M]_D$ (PLS)	$[M]_D$ (ANN)
47	b and 1,5-dihydroxy naphthalene	-24.90	-21.58	-23.42
48	c and 3,3'-diaminobenzophenone	-9.40	-6.84	-9.71
49	d and 1,3-phenylenediamine	-32.30	-37.40	-33.65
50	f and *p*-phenylenediamine	-9.70	-7.16	-9.50
51	g and 2,4-dihydroxyacetophenone	-6.38	-4.14	-6.17
52	h and 4,4'-diaminobiphenyl	-54.30	-43.15	-53.28
53	j and 4,4'-sulfonyldianiline	-26.60	-20.21	-27.46

a is *N,N′*-(4,4'-hexafluoroisopropylidendiphthaloyl)-bis-*L*-isoleucine

b is 4,4'-(hexafluoroisopropylidene)-*N,N′*-bis (phthaloyl-L-leucine) diacid chloride

c is 4,4'-(hexafluoroisopropylidene) bis (phthaloyl-L-leucine)

d is 4,4'–(hexafluoroisopropylidene)-*N,N′*-bis-(phthaloyl-L-methionine) diacid chloride

e is N,N'–(4,4(-hexafluoroisopropylidenediphthaloyl)-bis-L-methionine

f is *N,N* '-(4,4'-oxydiphthaloyl)-bis-*L*-isoleucine diacid chloride

g is *N,N′*-(4,4′-oxydiphthaloyl)-bis-*L*-leucine

h is *N,N′*-(4,4'-oxydiphthaloyl)-bis-L-methionine diacid chloride

i is *N,N′*-(4,4'-oxydiphthaloyl)-bis-(s)-(+)-valine diacid chloride

j is *N,N′*-(4,4'-carbonyldiphthaloyl)-bis-L-leucine diacid chloride

next step, the Hyperchem output files were used by the dragon package to calculate molecular descriptors. Dragon is new, freely available software (by Milano Chemometrics and the QSAR Research Group) for the calculation of more than 1400 molecular descriptors [35]. After the calculation of molecular descriptors, those that stayed constant and near constant for all molecules were removed from the descriptor pool, since those descriptors were not encoding the structural differences between compounds. Further reduction of the descriptor pool was attained by examining pairwise correlations between descriptors so that only one descriptor was retained from a pair contributing similar information (correlation coefficient >0.9 in this study). Finally, a total set of 648 remaining descriptors are achieved and used to select optimal subset of descriptors.

Descriptors selection and QSAR models development

GA-PLS variable selection

The strategy implemented for genetic algorithm-based variable selection in the frame of PLS regression can be described through the different steps detailed in ref. [36] GA-PLS is a sophisticated hybrid approach that combines GA [37] as a powerful optimization method with PLS [38–40] as a robust statistical method for variable selection. The combination of variables and the internal predictivity of the derived PLS model in GA-PLS correspond a chromosome and its fitness in GA, respectively. GA-PLS consists of three basic steps. (1) An initial population of chromosomes is created. Each chromosome is a binary bit string, by which the existence of a variable is represented. (2) A fitness of each chromosome in the population is evaluated by the internal predictivity of PLS. (3) The population of chromosomes in the next generation is reproduced. Three operations, i.e., selection, cross-over and mutation of chromosomes, are made in this step. In the overall scheme, steps 2 and 3 are continued until the number of the repetitions is reached at the designated number of generations.

In this paper, GA-PLS followed Leardi’s method [36]. The values of empirical parameters affecting the performance of GA-PLS were defined as in Table 2. Because each GA gives a slightly different model, repeat each run at least five times to verify the robustness of the predictive ability and importance of the selected model. If some variables (descriptors) are present only in one model, it can be concluded that they have selected by chance and therefore, they can be disregarded in the final model.

Partial least squares (PLS)

Partial least squares (PLS) regression is a modern technique that generalizes and combines features from principal component analysis and multiple regression. It is particularly helpful when we need to predict a set of dependent variables from a (very) large set of independent variables (i.e., predictors). PLS regression has acquired a famous position in chemometrics [41]. One reason for this is that it overcomes the deficiencies of ordinary least squares (OLS) regression in the case of highly collinear data. Besides, PLS allows an analysis of the data in terms of independent latent variables or components. These PLS components span a

Table 2 Parameters of the genetic algorithm

Population size	30 Chromosomes
Regression method	PLS
Maximum number of variables selected in the same chromosome	30
Maximum number of components	The optimal number
Response	Cross-validated % explained variance
Probability of mutation	0.1
Probability of cross over	0.5
Number of evaluation	200
Number of run	100

subspace of the regressors (columns of X) that is relevant for describing both X and the response Y. Ardent proponents of PLS consider it superior to other biased regression methods [42]. However, it is unlikely that there is a single superior technique for predictive modeling.

It is assumed that X (n×N) contains the descriptors that can be used for predicting the activities Y (n×M). It is distinguished that PLS decomposes the data matrices X and Y into a two matrices product plus residual in a single process. The matrices E and F contain residuals for X and Y, respectively:

$$X = TP' + E \tag{2}$$

$$Y = UQ' + F, \tag{3}$$

where T and U are score matrices and P'and Q' are loading matrices for X and Y, respectively. These two equations can be written as a multiple regression model:

$$Y = XB + G, \tag{4}$$

where matrix B contains the PLS regression coefficients [43].

The PLS algorithm used in this study was the singular value decomposition (SVD)-based PLS. This algorithm was proposed by Lobert et al. in 1987 [44]. A brief discussion of the SVD-based PLS algorithm can be found in the literature [45–47]. The program of PLS modeling based on SVD was written with MATLAB 7 in our laboratory [48].

Artificial neural network

The ANN is a computer-based system derived from a basic idea of the brain in which a number of nodes, called progressing elements or neurons, are interconnected in a network [49, 50]. A detailed description of the theory behind a neural network has been adequately described elsewhere [51–53]. There is a range of artificial neural network architectures designed and used in various fields. In this study, a feed-forward neural network with back propagation learning algorithm is used. The basic element of a back-propagation neural network is the processing node. Each processing node behaves like a biological neuron and performs two functions. First, it sums the values of its inputs. This sum is then passed through a transfer function to generate an output. Any differentiable function can be used as transfer function, *f*. All the processing nodes are arranged into layers, each fully interconnected to the following layer. There is no interconnection between the nodes of the same layer. In a back-propagation neural network, generally, there is an input layer that acts as a distribution structure for the data being presented to the network. This layer is not used for any type of processing. After this layer, one or more processing layers follow, called the hidden layers. The final processing layer is called the output layer.

In the present work, an ANN program was written with MATLAB 7. This network was feed-forward fully connected that has three layers with sigmoidal transfer function. Descriptors selected by GA and PLS methods were used as inputs of network and its output signal represent the optical rotation of interested macromolecules. Thus this network has six nodes in input layer and one node in output layer. The value of each input was divided into its mean value to bring them into dynamic range of the sigmoidal transfer function of the network. The initial values of weights were randomly selected from a uniform distribution that ranged between -0.3 to +0.3 and the initial values of biases were set to be one. These values were optimized during the network training. The back-propagation algorithm was used for the training of the network. Before training, the network parameters would be optimized. These parameters are: number of nodes in the hidden layer, weights and biases learning rates and the momentum. Procedures for the optimization of these parameters were reported elsewhere [54, 55]. Then the optimized network was trained using training set for adjustment of weights and biases values. To maintain the predictive power of the network at a desirable level, training was stopped when the value of error for the test set started to increase. Since the test error is not a good estimation of the generalization error, the prediction potential of the model was evaluated on a third set of data, named validation set. Compounds in the validation set were not used during the training process and were reserved to evaluate the predictive power of the generated ANN.

Estimation of the predictive ability of a QSAR model

For the optimized QSAR model several parameters were selected to test prediction ability of the model. A real QSAR model may have a high predictive ability, if it is close to ideal one. This may imply that the correlation coefficient R between the experimental (actual) y and predicted $\tilde{y}$ properties must be close to 1 and regression of y against $\tilde{y}$, i.,e. $y^{r0} = k\tilde{y}$ should be characterized by k close to 1 [56]. Slopes k is calculated as follows:

$$k = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2}. \tag{5}$$

The criteria formulated above may not be sufficient for a QSAR model to be truly predictive. Regression line through the origin defined by $y^{r0} = k\tilde{y}$ (with the intercept set to one) should be close to optimum regression line $y^r = a\tilde{y} + b$. Correlation coefficient for this line R_0^2 is calculated as follows:

$$R_0^2 = 1 - \frac{\sum \left(\tilde{y}_i - y_i^{r0}\right)^2}{\sum \left(\tilde{y}_i - \bar{\tilde{y}}\right)^2}, \tag{6}$$

where $\bar{\tilde{y}}$ is the average value of the observed property and the summations are over all n compounds in the validation set.

A difference between R^2 and R_0^2values (R_m^2) needs to be studied to explore the prediction potential of a model [57]. This term was defined in the following manner:

$$R_m^2 = R^2\left(1 - \left|\sqrt{R^2 - R_0^2}\right|\right). \tag{7}$$

Finally, the following criteria for evaluation of the predictive ability of QSAR models should be considered:

1. High value of cross-validated R^2 (q^2>0.5).
2. Correlation coefficient R between the predicted and actual properties from an external test set close to 1. R_0^2 should be close to R^2.
3. Slope of regression line (k) through the origin should be close to 1.
4. R_m^2 should be greater than 0.5.

Results and discussion

PLS modeling

Table 1 shows the data set and corresponding observed PLS and ANN predicted values of optical rotation of all polymers studied in this work. Parameters of genetic algorithm for generation of GA-PLS are shown in Table 2. Table 3 shows the specifications of best PLS model. The optimum number of latent variables to be included in the model was three. It can be seen from this table that six descriptors appeared in this model. These descriptors are: gravitational index (G1), average valence connectivity index chi-3 (X3AV), 3D-Harary index (H3D), 3st component symmetry directional WHIM index/weighted by atomic Van der Waals volumes (G3V), d CoMMA2 value/ weighted by atomic polarizabilities (DISPP), and mean topological charge index of order10 (JGI10). Each of these descriptors encodes different aspects of the molecular structure. The numerical values of these descriptors are shown in Table 4. Table 5 represents the correlation matrix for these descriptors. By interpreting the descriptors in this model, it is possible to gain some insight into factors that are likely related to the optical rotation of the polymers.

For inspection of the relative importance and contribution of each descriptor in the model, the value of mean effect (ME) was calculated for each descriptor by the following equation:

$$ME_j = \frac{\beta_j \sum_{i=1}^{n} d_{ij}}{\sum_j^m \beta_j \sum_i^n d_{ij}}, \tag{8}$$

where, ME_j is the mean effect for considered descriptor *j*, β_j

Table 3 Specification of partial least squares (PLS) method

Descriptor	Notation	Coefficient	Mean effect
Gravitational index	G1	-0.233	-38.941
Average valence connectivity index chi-3	X3AV	-910.651	-72.089
3D-Harary index	H3D	0.002	2.765
3st Component symmetry directional WHIM index /weighted by atomic Van der Waals volumes	G3V	-1586.036	-216.130
d CoMMA2 value / weighted by atomic polarizabilities	DISPP	-23.282	-9.895
Mean topological charge index of order10	JGI10	-1339.827	-16.802
Constant		330.887	

Table 4 The values of the descriptors that were used in this work[a]

Number	G1	X3AV	H3D	G3V	DISPP	JGI10
1	227.132	0.081	254.350	0.134	0.239	0.015
2	208.374	0.084	229.090	0.132	0.962	0.013
3	112.484	0.077	284.580	0.134	0.521	0.014
4	123.840	0.075	208.360	0.133	0.581	0.015
5	185.239	0.073	260.650	0.132	0.828	0.017
6	159.378	0.079	257.750	0.130	0.233	0.019
7	159.333	0.073	206.760	0.140	0.212	0.017
8	135.524	0.071	247.770	0.131	0.778	0.019
9	189.576	0.074	225.290	0.132	0.478	0.018
10	126.759	0.080	239.470	0.132	0.819	0.018
11	181.687	0.085	920.790	0.135	0.215	0.019
12	122.058	0.073	237.830	0.139	0.186	0.017
13	169.241	0.081	931.560	0.135	0.187	0.013
14	161.790	0.078	876.720	0.139	0.177	0.013
15	258.851	0.078	868.740	0.139	0.158	0.012
16	125.151	0.079	240.370	0.137	0.316	0.013
17	121.153	0.079	190.970	0.136	0.161	0.013
18	111.276	0.075	178.260	0.139	0.174	0.012
19	127.816	0.073	192.670	0.140	0.167	0.012
20	215.127	0.094	1029.640	0.137	0.603	0.012
21	238.098	0.085	1033.340	0.137	0.539	0.010
22	281.334	0.082	11026.920	0.133	0.551	0.010
23	125.824	0.083	955.540	0.137	0.554	0.008
24	199.757	0.086	1029.810	0.141	0.523	0.010
25	155.780	0.075	348.360	0.130	0.807	0.011
26	82.939	0.071	288.510	0.143	0.588	0.009
27	93.847	0.078	283.260	0.138	0.636	0.008
28	91.363	0.078	303.300	0.136	0.625	0.009
29	246.203	0.099	1002.690	0.143	0.308	0.011
30	431.840	0.088	2006.470	0.122	0.178	0.010
31	218.947	0.079	1014.480	0.144	0.450	0.010
32	247.303	0.075	11008.700	0.135	0.429	0.010
33	162.615	0.075	945.690	0.139	0.429	0.009
34	167.999	0.075	938.050	0.139	0.428	0.008
35	118.058	0.082	281.590	0.135	0.216	0.010
36	145.376	0.078	268.210	0.142	0.234	0.010
37	232.027	0.078	10330.490	0.142	0.236	0.010
38	173.134	0.077	203.840	0.140	0.219	0.013
39	175.602	0.071	225.260	0.131	0.359	0.019
40	129.203	0.075	945.830	0.136	0.223	0.018
41	287.820	0.078	1095.200	0.131	0.239	0.013
42	112.827	0.073	205.730	0.135	0.669	0.012
43	96.921	0.075	313.780	0.134	0.877	0.010
44	364.954	0.088	1979.950	0.124	0.232	0.010
45	175.767	0.077	1011.700	0.135	0.476	0.011
46	144.741	0.078	192.850	0.139	0.200	0.013
47	173.501	0.073	246.080	0.133	0.530	0.017
48	152.610	0.075	921.140	0.131	0.222	0.017
49	249.474	0.079	945.510	0.136	0.243	0.014

Table 4 (continued)

Number	G1	X3AV	H3D	G3V	DISPP	JGI10
50	154.444	0.073	963.200	0.134	0.555	0.009
51	93.133	0.076	300.250	0.136	0.556	0.012
52	210.846	0.088	1010.690	0.139	0.498	0.011
53	144.968	0.090	309.710	0.135	0.191	0.013

The definitions of the descriptors are given in Table 3

[a] The numbers refer to the numbers of the molecules given in Table 1

is the coefficient of descriptor j and d_{ij} is the value of interested descriptors for each molecule, and m is the number of descriptors in the model. The calculated values of MEs are represented in the last column of Table 3 and are also plotted in Fig. 2. The value and sign of mean effect shows the relative contribution and direction of influence of each descriptor on the optical rotation.

Of the six descriptors, three are geometrical (G1, H3D and DISPP), one is topological (X3AV), one is WHIM (G3V) and one is Galves topological charge index (JGI10). G1 reflects the mass distribution in a molecule and defined as:

$$G_1 = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} \frac{m_i m_j}{r_{ij}^2}, \qquad (9)$$

where m$_i$, and m$_j$, are the atomic masses of the considered atoms, r$_{ij}$ is the corresponding interatomic distances and A is the number of atoms of the molecule, respectively. The G$_1$ index takes into account all atom pairs in the molecule. This index is related to the bulk cohesiveness of the molecules, accounting, simultaneously, for both atomic masses (volumes) and their distribution within the molecular space [58]. Harary index (also called Harary number) is a molecular topological index [59, 60] derived from the reciprocal distance matrix D^{-1} by the Wiener operator W:

$$H = W(D)^{-1} = \sum_{i=1}^{A} \sum_{j=1}^{A} d_{ij}^{-1}. \qquad (10)$$

The Harary index increases with both molecular size and molecular branching; it is therefore a measure of molecular compactness. DISPP is comparative molecular moment analysis (CoMMA) descriptor. Geometrical representation of the molecule calculates different molecular moments with respect to the center of mass, center of charge and center of dipole of the molecule [61, 62]. By calculating molecular descriptors based on 3D geometry without a common orientation frame, the CoMMA overcomes the problems due to the molecular alignment. X3AV is a valence connectivity index and can be calculated as follows:

$$^{m}\chi_{AV} = \sum_{k=1}^{k} \left(\prod_{a=1}^{n} \delta^{v} \right)_k, \qquad (11)$$

where k runs over all of the m^{th} order subgraphs constituted by n atoms (n = m + 1 for acyclic subgraphs); K is the total number of m^{th} order subgraphs present in the molecular graph and in the case of the path subgraphs equals the m^{th} order path count and $\delta^v$is valence vertex degree. This topological descriptor (also called topological index) describes the atomic connectivity in the molecule [63, 64]. JGI10, a topological charge index was proposed to evaluate the charge transfer between pairs of atoms, and therefore the global charge transfer in the molecule [65, 66]. As can be seen in Table 3, from the above mentioned descriptors

Table 5 Correlation matrix between selected descriptors

	G1	X3AV	H3D	G3V	DISPP	JGI10
G1	1	0.494	0.451	-0.317	-0.146	-0.126
X3AV		1	0.043	-0.026	-0.055	-0.205
H3D			1	0.011	-0.048	-0.276
G3V				1	-0.277	-0.378
DISPP					1	0.003
JGI10						1

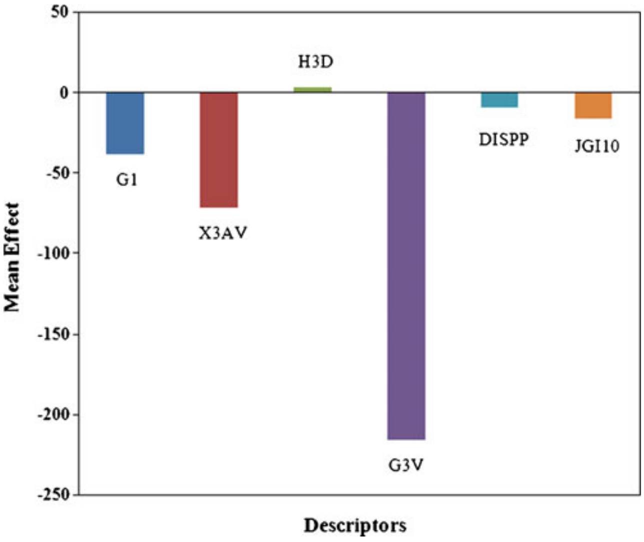

Fig. 2 Plot of descriptor's mean effects

Table 6 Architecture and specifications of optimized ANN model

Number of nodes in the input layer	6
Number of nodes in the hidden layer	5
Number of nodes in the output layer	1
Weights learning rate	0.3
Biases learning rate	0.2
Momentum	0.4
Transfer function	Sigmoid

only H3D has a positive sign for its mean effect. This means that increasing the value of this descriptor causes the increasing of the values of optical rotations and increasing the values of the other descriptors decreases the $[M]_D$ values.

From the above discussion, it can be seen that all descriptors involved in the QSAR model have physical meaning, and these descriptors can account for structural features that affect the optical rotation of the interested molecules.

Neural network modeling

The next step was the construction of an ANN. During the training of the ANNs, the parameters of network including the number of nodes in the hidden layer, weights and biases learning rates and momentum values were optimized. Table 6 shows the architecture and specification of the optimized network. After optimization of the network parameters, the network was trained by using training set for adjustment of the weights and biases values by back-propagation algorithm. It is known that a neural network can become over-trained. An over-trained network has usually learned perfectly the stimulus pattern it has seen but can not give accurate prediction for unseen stimuli. There are several methods for overcoming this problem. One method is to use a test set to evaluate the prediction power of the network during its training. In this method after each 1000 training iterations, the network was used to calculate $[M]_D$ of molecules included in the test set. To maintain the predictive power of the network at a desirable level, training was stopped when the value of errors for the test set started to increase. Results obtained showed over-training began after 48000 iterations.

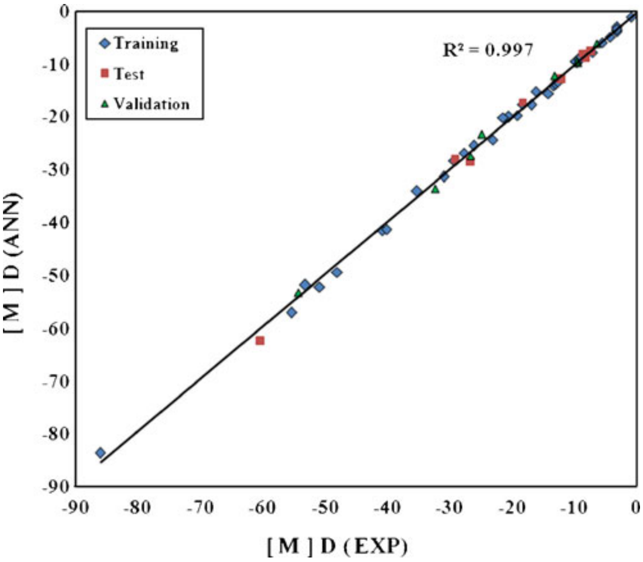

Fig. 3 Plot of ANN calculated optical rotation against experimental values

The predictive power of the ANN models developed on the selected training sets are estimated on the predictions of validation set chemicals, by calculating the q^2 that is defined as follow:

$$q^2 = 1 - \frac{\sum (y_i - \widehat{y}_i)^2}{\sum (y_i - \bar{y})^2}, \tag{12}$$

where y_i and $\widehat{y}_i$, respectively are the measured and predicted values of the dependent variable (optical rotation), $\bar{y}$ is the averaged value of dependent variable of the training set and the summations cover all the compounds. The calculated value of q^2 was 0.996.

Table 1 represents the experimental, PLS and ANN calculated values of optical rotation for the training, test and validation sets. The statistical parameters obtained by ANN and PLS models for these sets are shown in Table 7. The standard errors of training, test and validation sets for the PLS model are 5.494, 6.519, and 5.331, respectively which would be compared with the values of 0.925, 1.132, and 1.047, respectively, for the ANN model. Comparison between these values and other statistical parameters in Table 7 reveals the superiority of the ANN model over PLS one. The key strength of neural networks, unlike PLS analysis, is their ability to do flexible mapping of the

Table 7 Statistical parameters obtained using the ANN and PLS models[a]

Model	SE_c	SE_t	SE_v	R_c^2	R_t^2	R_v^2	F_c	F_t	F_v
ANN	0.925	1.132	1.047	0.998	0.996	0.996	14688	2226	1645
PLS	5.494	6.519	5.331	0.916	0.885	0.906	383	47	58

[a] c refers to the calibration (training) set; t refers to test set; v refers to validation set; R is the correlation coefficient; SE is standard error and F is the statistical F value

selected features by manipulating their functional dependence implicitly.

The statistical values of validation set for the ANN model was characterized by q2 =0.996, R^2 =0.996, $R_0^2 = 0.996$, $R_m^2 = 0.981$ and k=1.006. These values and other statistical parameters which are shown in Table 7 reveal the high predictive ability of the model. Figure 3 shows the plot of the ANN predicted versus experimental values for optical rotation of all of the molecules in data set.

Conclusions

In the present work GA as a feature selection tool and PLS and ANN as feature mapping techniques were used for prediction of the optical rotation of 53 biodegradable polymers. The optimized 6-5-1 ANN model showed a remarkable improvement over the linear model. The GA-based PLS approach is especially useful for modeling a large variable data set. The physical meaning of the selected subset of descriptors, which are the most predictive and informative, from the GA method, is determined. The optical rotations of investigated polymers were interpreted rationally with these six descriptors. The squared correlation coefficient, R^2 values of the PLS model for the training, test and validation sets were 0.916, 0.885 and 0.906 respectively which would be compared with the values of 0.998, 0.996 and 0.996, respectively, for the ANN model. Results obtained indicate that while the GA and PLS methods could be more powerful in precise selecting of important parameters and assume the significance of each of descriptors, introduction of neural network gives a significant improvement of prediction quality.

Acknowledgments We wish to express our gratitude to the Research Affairs Division Isfahan University of Technology (IUT), Isfahan, for partial financial support. Further financial support from National Elite Foundation (NEF) and Center of Excellency in Sensors and Green Chemistry Research (IUT) is gratefully acknowledged.

References

1. Barron LD (2004) Molecular light scattering and optical activity, 2nd edn. Cambridge University Press, Cambridge, UK
2. Charney E (1979) The molecular basis of optical activity: optical rotatory dispersion and circular dichroism. Wiley, New York
3. Miguela RN, Sastrea JAL, Galisteoa D, Martına AD, Ramos AG (2000) Calculation of optical rotation from molecular structure: comparative study of MM2, MM3 and AM1 methods. J Mol Struct 522:219–231
4. Ruud K, Helgaker T (2002) Optical rotation studied by density-functional and coupled-cluster methods. Chem Phys Lett 352:533–539
5. Stephens PJ, Devlin FJ, Cheeseman JR, Frisch MJ, Mennuccic B, Tomasic S (2000) Synthesis of optically active a-methylene g-lactones through lipase-catalyzed kinetic resolution. Tetrahedron: Asymmetry 11:2443–2448
6. Xinliang Y, Zhimin X, Bing Y, Xueye W, Fang L (2007) Prediction of the thermal decomposition property of polymers using quantum chemical descriptors. Eur Polym J 43:818–823
7. Xiaobing L, Shijun L, Jinhua P, Xueye W (2009) Theoretical study on sulfonated and phosphonated poly[(aryloxy)phosphazenes] as proton-conducting membranes for fuel cell applications. Eur Polym J 45:2391–2394
8. Aihong L, Xueye W, Ling W, Hanlu W, Hengliang W (2007) Prediction of dielectric constants and glass transition temperatures of polymers by quantitative structure property relationships. Eur Polym J 43:989–995
9. Mallakpour S, Hatami M, Golmohammadi H (2010) Prediction of inherent viscosity for polymers containing natural amino acids from the theoretical derived molecular descriptors. Polymer 51:3568–3574
10. Polavarapu PL, Chakraborty DK, Ruud K (2000) Molecular optical rotation: an evaluation of semiempirical models. Chem Phys Lett 319:595–600
11. Amos RD (1982) Electric and magnetic properties of CO, HF, HCI, and CH_3F. Chem Phys Lett 87:23–26
12. Helgaker T, Ruud K, Bak KL, Jørgensen P, Olsen J (1994) Vibrational Raman optical activity calculations using London atomic orbitals. Faraday Discuss 99:165–180
13. Polavarapu PL (1997) Vibrational optical activity of anharmonic oscillator. Mol Phys 91:551–554
14. Stephens PJ, Devlin FJ, Cheeseman JR, Frisch MJ (2001) Calculation of Optical Rotation Using Density Functional Theory. J Phys Chem A 105:356–367
15. Yao X, Zhang X, Zhang R, Liu M, Hu Z, Fan B (2001) Prediction of enthalpy of alkanes by the use of radial basis function neural networks. Computers and Chemistry 25:475–482
16. Schweitzer RC, Morris JB (1999) The development of a quantitative structure property relationship (QSPR) for the prediction of dielectric constants using neural networks. Anal Chem Acta 384:285–303
17. Fatemi MH (2002) Simultaneous modeling of the Kovats retention indices on OV-1 and SE-54 stationary phases using artificial neural networks. J Chromatogr A 955:273–280
18. Golmohammadi H, Fatemi MH (2005) Artificial neural network prediction of retention factors of some benzene derivatives and heterocyclic compounds in micellar electrokinetic chromatography. Electrophoresis 26:3438–3444
19. Baher E, Fatemi MH, Konoz E, Golmohammadi H (2007) Prediction of retention factors in micellar electrokinetic chromatography from theoretically derived molecular descriptors. Microchim Acta 158:117–122
20. Konoz E, Golmohammadi H (2008) Prediction of air-to-blood partition coefficients of volatile organic compounds using genetic algorithm and artificial neural network. Anal Chem Act 619:157–164
21. Golmohammadi H (2009) Prediction of octanol–water partition coefficients of organic compounds by multiple linear regression, partial least squares, and artificial neural network. J Comput Chem 30:2455–2465
22. Golmohammadi H, Konoz E, Dashtbozorgi Z (2009) Prediction of gas-to-olive oil partition coefficients of organic compounds using an artificial neural network. Anal Sci 25:1137–1142
23. Mallakpour SE, Hajipour A, Khoee S (2002) Rapid synthesis of optically active poly(amide–imide)s by direct polycondensation of aromatic dicarboxylic acid with aromatic diamines. Eur Polym J 38:2011–2016
24. Mallakpour S, Moghaddam E (2006) Preparation of new poly(ester-imide)s from n, n'-(4, 4'-hexafluoroisopropylidendiphthaloyl)-bis-l-isoleucine and aromatic diols with tscl/py/dmf as a condensing agent. Iran Polym J 15:547–554

25. Mallakpour SE, Hajipour A, Khoee S (2000) Microwave-Assisted Polycondensation of 4, 4'-(Hexafluoroisopropylidene)-N, N'-bis (phthaloyl-L-leucine) Diacid Chloride with Aromatic Diols. J Appl Poly Sci 77:3003–3009
26. Mallakpour SE, Hajipour A, Khoee S (1999) Synthesis and characterization of novel optically active poly(amide-imide)s. Polym Int 48:1133–1140
27. Mallakpour S, Kowsari E (2006) Thermally stable and optically active poly(amideimide)s derived from 4, 4'–(hexafluoroisopropylidene)-n, n'-bis-(phthaloyl-l-methionine) diacid chloride and various aromatic diamines: synthesis and characterization. Polym Bull 57:169–178
28. Mallakpour S, Kowsari E (2006) Preparation and characterization of new thermally stable and optically active poly(ester-imide)s by direct polycondensation with thionyl chloride in pyridine. Polym Adv Technol 17:174–179
29. Mallakpour S, Kowsari E (2005) Polycondensation reaction of n, n'-(4, 4'-oxydiphthaloyl)-bis-l-isoleucine diacid chloride with aromatic diamines. Iran Polym J 14:799–806
30. Mallakpour S, Kowsari E (2006) Thionyl chloride/pyridine system as a condensing agent for the polyesterification reaction of n, n′-(4, 4′-oxydiphthaloyl)-bis-l-leucine and aromatic diols. Iran Polym J 15:457–465
31. Mallakpour S, Habibi S (2003) Microwave-promoted synthesis of new optically active poly(ester-imide)s derived from N, N'-(pyromellitoyl)-bis-L-leucine diacid chloride and aromatic diols. Eur Polym J 39:1823–1829
32. Mallakpour SE, Hajipour A, Zamanlou MR (2001) Synthesis of optically active poly(amide-imide)s derived from n, n'-(4, 4'-carbonyldiphthaloyl)-bis-l-leucine diacid chloride and aromatic diamines by microwave radiation. J Polym Sci 39:177–186
33. Hyperchem, re. 4. for Windows, Autodesk, Sansalito, CA, 1995
34. Mopac for Windows, Stewart Computational Chemistry, 2009
35. Mauri A, Consonni V, Pavan M, Todeschini R (2006) DRAGON software: An easy approach to molecular descriptor calculations. MATCH Commun Math Comput Chem 56:237–248
36. Leardi R, Gonzales AL (1998) Genetic algorithms applied to feature selection in PLS regression: how and when to use them. Chemom Intell Lab Syst 41:195–207
37. Goldberg DE (1989) Genetic algorithms in search. Optimization and Machine learning, Addison-Wesley, New York
38. Kowalski B, Gerlach R (1982) In: Joreskog KG, Wold H (eds) Systems under indirect observatio. North Holland, Amsterdam, pp 191–209
39. Yu RQ (1992) Introduction to chemometrics. Human Education, Changsha
40. Dayal BS, MacGregor JF (1997) Improved PLS algorithms. J Chemom 11:73–85
41. Martens H, Næs T (1992) Multivariate calibration. Wiley, Chichester
42. Hoskuldsson A (1992) The H-principle in modelling with applications to chemometrics. Chemom Intell Lab Syst 14:139–153
43. Wold S, Sjostorm M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58:109–130
44. Lorber A, Wangen L, Kowalsky BR (1987) A theoretical foundation for the PLS algorithm. J Chemom 1:19–31
45. Khayamian T, Ensafi AA, Hemmateenejad B (1999) Simultaneous spectrophotometric determinations of cobalt, nickel and copper using partial least squares based on singular value decomposition. Talanta 49:587–596
46. Shamsipur M, Hemmateenejad B, Akhond M, Sharghi H (2001) Quantitative structure–property relationship study of acidity constants of some 9, 10-anthraquinone derivatives using multiple linear regression and partial least-squares procedures. Talanta 54:1113–1120
47. Hoskuldsson A (2001) Variable and subset selection in PLS regression. Chemom Intell Lab Syst 55:23–38
48. MATLAB 7.0, The Mathworks Inc., Natick, MA, USA, http://www.mathworks.com
49. Husain S, Devi KS, Krishna D, Reddy PJ (1996) Characterization and identification of edible oil blends and prediction of the composition by artificial neural networks - a case study. Chemom Intell Lab Syst 35:117–126
50. Holland JH (1992) Adaption in neural and artificial systems. MIT Press, MA, Cambridge
51. Zupan J, Gasteiger J (1999) Neural network in chemistry and drug design. Wiley-VCH, Weinheim
52. Beal TM, Hagan HB, Demuth M (1996) Neural Network Design; PWS, Boston
53. Zupan J, Gasteiger J (1993) Neural networks for chemists: an introduction. Weinheim, VCH
54. Blank TB, Brown ST (1993) Nonlinear multivariate mapping of chemical data using feed-forward neural networks. Anal Chem 65:3081–3089
55. Jalali-Heravi M, Fatemi MH (2001) Artificial neural network modeling of Kova´ts retention indices for noncyclic and monocyclic terpenes. J Chromatogr A 915:177–183
56. Golbraikh A, Tropsha A (2002) Beware of q^2! J Mol Graphics Modell 20:269–276
57. Roy PP, Roy K (2008) On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci 27:302–313
58. Katritzky AR, Mu L, Lobanov VS, Karelson M (1996) Correlation of boiling points with molecular structure. 1. A training set of 298 diverse organics and a test set of 9 simple inorganics. J Phys Chem 100:10400–10407
59. Ivanciuc O, Balaban TS, Balaban AT (1993) Design of topological indices. Part 4. Reciprocal distance matrix, related local vertex invariants and topological indices. J Math Chem 12:309–318
60. Plavsic D, Nikolic S, Trinajstic N, Mihalic Z (1993) On the Harary index for the characterization of chemical graphs. J Math Chem 12:235–250
61. Platt DE, Silverman BD (1996) Registration, orientation, and similarity of molecular electrostatic potentials through multipole matching. J Comput Chem 17:358–366
62. Silverman BD, Platt DE (1996) Comparative molecular moment analysis (comma): 3d-qsar without molecular superposition. J Med Chem 39:2129–2140
63. Kier LB, Hall LH (1981) Derivation and significance of valence molecular connectivity. J Pharm Sci 70:583–589
64. Kier LB, Hall LH (1983) General definition of valence delta-values for molecular connectivity. J Pharm Sci 72:1170–1173
65. Gilvez J, Garcia R, Salabert MT, Soler R (1994) Ab initiu molecular optical rotations and absolute configurations. J Chem Inf Comput Sci 34:520–525
66. Gilvez J, Garcia-Domenech R, De Juliin-Ortiz V, Soler R (1995) Topological approach to drug design. J Chem Inf Comput Sci 35:272–284