Journal of Computer-Aided Molecular Design (2005) 19: 509–524
DOI 10.1007/s10822-005-9004-4

Quantitative prediction of imprinting factor of molecularly imprinted polymers by artificial neural network

Chanin Nantasenamat[a], Thanakorn Naenna[b], Chartchalerm Isarankura Na Ayudhya[a] & Virapong Prachayasittikul[a,*]
[a]*Department of Clinical Microbiology, Faculty of Medical Technology, Mahidol University, 10700, Bangkok, Thailand;* [b]*Department of Industrial Engineering, Faculty of Engineering, Mahidol University, 73170, Nakhon Pathom, Thailand*

Received 22 April 2005; accepted 11 July 2005

Key words: artificial neural network, back-propagation, data mining, molecular imprinting, molecularly imprinted polymer

Summary

Artificial neural network (ANN) implementing the back-propagation algorithm was applied for the calculation of the imprinting factors (IF) of molecularly imprinted polymers (MIP) as a function of the computed molecular descriptors of template and functional monomer molecules and mobile phase descriptors. The dataset used in our study were obtained from the literature and classified into two distinctive datasets on the basis of the polymer's morphology, irregularly sized MIP and uniformly sized MIP datasets. Results revealed that artificial neural network was able to perform well on datasets derived from uniformly sized MIP ($n = 23$, $r = 0.946$, RMS = 2.944) while performing poorly on datasets derived from irregularly sized MIP ($n = 75$, $r = 0.382$, RMS = 6.123). The superior performance of the uniformly sized MIP dataset over the irregularly sized MIP dataset could be attributed to its more predictable nature owing to the consistency of MIP particles, uniform number and association constant of binding sites, and minimal deviation of the imprinted polymers. The ability to predict the imprinting factor of imprinted polymer prior to performing actual experimental work provide great insights on the feasibility of the interaction between template-functional monomer pairs.

Abbreviations: MIP – molecularly imprinted polymer; NIP – non-imprinted polymer; IF – imprinting factor; RMS – root mean square error

Introduction

Molecular imprinting is a promising technology that confers specific molecular recognition capabilities to synthetic polymers [1]. The molecular imprinting process involves the formation of a complex between a target (template) molecule with appropriate functional monomer(s) that cross-links into a macromolecular matrix. Subsequent removal of the template reveals binding sites that are complementary in size and shape to the original imprint molecule. Molecularly imprinted polymers (MIP) have been prominently used as separation media both in chromatography and solid phase extraction achieving enantioselective separation of racemic mixtures of compounds [2]. Moreover, MIP has also been applied as recognition elements in biological and chemical sensors [3, 4], mimics of antibodies and receptors used in drug assays [5], mimics of enzymes [6], and production factories for small enzyme inhibitors [7].

*To whom correspondence should be addressed. Fax: +662-849-6330; E-mail: mtvpr@mahidol.ac.th

In attempts to improve the recognition abilities of MIP to be highly selective and specific to target molecules, many research groups have focused on ways to obtain stable template-functional monomer interaction as it is the basis of MIP formation. Sellergren and coworkers utilized ^{1}H-NMR to detect adducts formation between the template and functional monomer [8]. Likewise, Duffy et al. employed infrared spectroscopy to monitor hydrogen bond formation [9]. Nicholls et al. stressed the importance of monomer-template interactions in governing the quality of the MIP binding sites [10].

In order to investigate the monomer-template interaction through computational approach, Piletsky and co-workers employed molecular modeling in the search of optimal functional monomer for a given template molecule [11–13]. Wu and coworkers established a correlation between the experimental retention factor (k') and the calculated interaction energy obtained from quantum mechanics/molecular mechanics (QM/MM) software [14, 15].

In our study, we propose the application of artificial neural networks in the field of molecular imprinting as a useful, economical, and time-saving technique in elucidating the feasibility of potential template-functional monomer pairs prior to performing actual experiments. This was carried out by collecting data from the literature and predicting the imprinting factors using the calculated molecular and mobile phase descriptors as inputs. The feasibility of the template-functional monomer pair could be inferred from the imprinting factor, in that a strong interaction would yield a large imprinting factor, while weak interactions would give small imprinting factor. The dataset in our study could be classified into two distinctive groups, irregularly sized MIPs that are produced by the traditional bulk polymerization and uniformly sized MIPs produced by either multi-step swelling or aqueous suspension polymerization. The erratic nature of MIPs prepared by bulk polymerization, in particularly the inconsistent size and heterogeneity in the number and association constant of binding sites, produced eccentric results making it hard for the artificial neural network to predict the imprinting factors. On the contrary, MIPs prepared by multi-step swelling and aqueous suspension polymerization possessed a rather consistent nature and so gave high predictive power.

Materials and methods

Overview of molecular imprinting process

The production of molecularly imprinted polymers comprises of three main steps, namely the formation of covalent or non-covalent adducts between template and functional monomer molecules, their polymerization, and template removal from the synthesized polymer. Polymers that are prepared by the traditional bulk polymerization are obtained as monolithic blocks and rely on the grinding step to break the polymeric block into smaller irregularly sized particles. Next, they are filtered into a narrow size distribution by sieving. Alternative polymerization methods include aqueous suspension polymerization and multi-step swelling and polymerization to obtain MIP particles of uniform size and shape; please see [16] for more detailed explanation. Subsequently, the MIP particles are packed into chromatographic columns where they are ready to be analyzed by high-performance liquid chromatography (HPLC).

Data collection

Data were collected from the literature by extracting the following information: template, functional monomer, molecularly imprinted polymer (MIP) retention factor, non-imprinted polymer (NIP) retention factor, imprinting factor (IF), mobile phase composition, and the pH of aqueous buffers.

The imprinting factor was used as a measure of the strength of interaction between the template and its corresponding functional monomer and was calculated according to the following equation:

$$\mathrm{IF} = \frac{\mathrm{MIP}(k')}{\mathrm{NIP}(k')} \tag{1}$$

where MIP(k') is the retention factor of the molecularly imprinted polymer, while NIP(k') is the retention factor of the non-imprinted polymer.

The datasets were divided into two main groups: the uniformly sized MIP dataset and the irregularly sized MIP dataset. The former were produced by multi-step swelling and aqueous suspension polymerization yielding MIP particles with consistent uniformly sized MIP particles, while the latter are generated by the conventional

bulk polymerization which gives rather irregularly sized MIP particles due to the grinding process.

Descriptors generation

The templates and functional monomers were converted into two-dimensional structures with ChemAxon's MARVIN and saved in SMILES notation [17], which then serves as input for the generation of 248 transferable atom equivalent (TAE) molecular descriptors by RECON version 5.5. The TAE methodology used in RECON is based on Bader's quantum theory of atoms in molecules and was developed by Breneman and co-workers to rapidly generate "molecular charge densities and charge density-based electronic properties of molecules, using atomic charge density fragments precomputed from *ab initio* wavefunctions" [18, 19] The TAE molecular descriptors are suitable to serve as input variables for the prediction of imprinting factors of imprinted polymers because they account for the electronic properties of molecules, which are relevant in studies of molecular interactions [20].

The mobile phase descriptors for the uniformly sized MIP dataset include aqueous buffer pH, dielectric constant of binary solvents, and ionic strength of aqueous buffers (Table 1). The irregularly sized MIP dataset had dielectric constant of organic solvents as the only descriptor since no aqueous buffer was present in its mobile phase composition (Table 2).

Dielectric constant values were taken from the CRC Handbook of Chemistry and Physics [21]. The dielectric constants of binary solvents [22–24] was calculated with the following equation:

$$\varepsilon_m = \phi_1 \varepsilon_1 + \phi_2 \varepsilon_2 \tag{2}$$

where ε_m, ε_1, and ε_2 are the dielectric constants of the mixture and solvents 1 and 2, respectively, while ϕ_1 and ϕ_2 represent the volume fractions of solvents 1 and 2 in the mixture.

The ionic strength (I) of aqueous buffers [25] was calculated with the following equation:

$$\mathrm{I} = \frac{1}{2}\sum_i c_i z_i^2 \tag{3}$$

where c_i and z_i represents the concentration and charge of ion i in solution. The summations are performed for all the possible ions in solution.

Descriptors reduction

The dataset were normalized using WEKA so that values fall in the range of 0 to 1. The dataset were scaled to a range between 0 and 1 with the min–max normalization equation as follows:

$$x_{\mathrm{norm}} = \frac{x_i - x_{\mathrm{min}}}{x_{\mathrm{max}} - x_{\mathrm{min}}} \tag{4}$$

where x_{norm}, x_i, x_{min}, and x_{max} is the normalized data, the value of each instance, the minimum value, and the maximum value of the dataset, respectively.

In order to improve prediction performance and reduce computational calculation time, redundant and multi-collinear TAE molecular descriptors were removed from the dataset with UFS 1.8, which is a program based on the variable reduction algorithm called Unsupervised Forward Selection (UFS) [26]. Thus, the generated 248 TAE molecular descriptors of the template and functional monomer molecules were reduced to a range of 2 to 19 TAE descriptors for the uniformly sized MIP dataset and 2 to 21 TAE descriptors for the irregularly sized MIP dataset. Briefly, UFS removes variables with standard deviation less than the pre-defined *sdevmin*, while terminating when the squared multiple correlation coefficients of the remaining variables display values greater than the user adjustable R^2_{max}. The standard deviation was left as default at 0.0005 while the *r*-squared-max (R^2_{max}) was varied between 0 and 0.99. The optimal number of TAE molecular descriptors to use was those that exhibit low RMS, as shown in Figure 1.

Overview of artificial neural network architecture

A three-layer feed-forward neural network utilizing the back-propagation of error algorithm was used to model the imprinting factor of molecularly imprinted polymers as a function of molecular descriptors and mobile phase compositions. Essentially, a majority of neural networks are comprised of three layers, namely the input, hidden, and output layer. The input layer receives input data, the hidden layer performs processing and transformation of the input data, and the output layer relays the final results. Each layer contains neurons (nodes) and the numbers of nodes presented in the input and output layer depends on the

Table 1. The uniformly-sized MIP dataset obtained from the literature.

No.	Template	Functional monomer	MIP k'	NIP k'	IF	Mobile phase composition	Polymerization method	Source
1	p-t-Octylphenol	4-VP	10.40	7.32	1.42	Acetonitrile/PBS (pH 5.1; 20 mM) (50/50, v/v)	P1	S1
2	p-t-Butylphenol	4-VP	4.52	3.16	1.43	Acetonitrile/PBS (pH 5.1; 20 mM) (50/50, v/v)	P1	S1
3	(S)-Nilvadipine	MAA	10.10	6.53	1.55	Acetonitrile/Phosphate buffer (pH 6; 20 mM) (40/60, v/v)	P1	S2
4	(S)-Nilvadipine	TFMAA	11.00	5.82	1.89	Acetonitrile/Phosphate buffer (pH 6; 20 mM) (40/60, v/v)	P1	S2
5	(S)-Nilvadipine	2-VP	27.60	17.50	1.58	Acetonitrile/Phosphate buffer (pH 6; 20 mM) (40/60, v/v)	P1	S2
6	(S)-Nilvadipine	4-VP	24.50	12.60	1.94	Acetonitrile/Phosphate buffer (pH 6; 20 mM) (40/60, v/v)	P1	S2
7	Clenbuterol	MAA	13.02	1.84	7.08	Acetonitrile/PBS (pH 3.4; 10 mM) (85/15, v/v)	P1	S3
8	Clenbuterol	AAm	4.18	0.90	4.64	Acetonitrile/PBS (pH 2; 10 mM) (50/50, v/v)	P1	S3
9	Trimethoprim	MAA	22.00	8.30	2.65	Methanol/Acetate buffer, Sodium chloride (pH 5; 50 mM) (90/10, v/v)	P2	S4
10	Matrine	MAA	12.15	7.04	1.73	Methanol/Water (50/50, v/v)	P2	S5
11	Oxymatrine	MAA	6.18	4.66	1.33	Methanol/Water (50/50, v/v)	P2	S5
12	Piritrexim	MAA	18.50	8.15	2.27	Methanol/Acetate buffer (pH 4; 20 mM) (60/40, v/v)	P2	S6
13	Bisphenol A	MAA	2.77	1.31	2.11	Acetonitrile/PBS (pH 5.1; 20 mM) (50/50, v/v)	P1	S7
14	Bisphenol A	2-DAEMA	5.44	3.31	1.64	Acetonitrile/PBS (pH 5.1; 20 mM) (50/50, v/v)	P1	S7
15	Bisphenol A	4-VP	15.30	2.10	7.29	Acetonitrile/PBS (pH 6.1; 20 mM) (60/40, v/v)	P1	S7
16	β-estradiol	4-VP	11.20	4.76	2.35	Acetonitrile/PBS (pH 5.1; 20 mM) (50/50, v/v)	P1	S7
17	(S)-Propanolol	MAA	35.90	8.52	4.21	Acetonitrile/PBS (pH 5.1; 20 mM) (70/30, v/v)	P1	S8
18	Cinchonine	MAA	54.10	1.71	31.64	Acetonitrile/Phosphoric acid, Sodium phosphate (pH 5.9; 50 mM) (50/50, v/v)	P1	S9
19	Cinchonidine	MAA	61.50	1.71	35.96	Acetonitrile/Phosphoric acid, Sodium phosphate (pH 5.9; 50 mM) (50/50, v/v)	P1	S9

20	(S)-Naproxen	4-VP	8.35	2.11	3.96	Acetonitrile/Phosphate buffer (pH 5.08; 20 mM) (50/50, v/v)	P1	S10
21	(S)-Ibuprofen	4-VP	4.42	2.24	1.97	Acetonitrile/Phosphate buffer (pH 5.08; 20 mM) (50/50, v/v)	P1	S10
22	d-Chlorpheniramine	MAA	19.40	3.00	6.47	Acetonitrile/Phosphoric acid, Potassium phosphate (pH 6.2; 50 mM) (70/30, v/v)	P1	S11
23	d-Chlorpheniramine	TFMAA	70.60	5.00	14.12	Acetonitrile/Phosphoric acid, Potassium phosphate (pH 5.3; 50 mM) (70/30, v/v)	P1	S11

MIP k', Retention factor of molecularly imprinted polymer; NIP k', Retention factor of non-imprinted polymer; IF, Imprinting factor; 2-DAEMA, 2-(diethylamino)ethyl methacrylate; 2-TFMAA, 2-(trifluoromethyl)acrylic acid; 2-VP, 2-vinylpyridine; 4-VP, 4-vinylpyridine; AAm, acrylamide; MAA, methacrylic acid.
P1, Multi-step swelling and polymerization; P2, Aqueous micro-suspension polymerization.
S1, Haruyo et al., Anal. Sci. 19, 715–719, 2003; S2, Fu et al., Anal. Chem. 75, 191–198, 2003; S3, Masci et al., J. Appl. Polym. Sci. 83, 2660–2668, 2002; S4, Lai et al., Anal. Bioanal. Chem. 372, 391–396, 2002; S5, Lai et al., Anal. Bioanal. Chem. 375, 264–269, 2003; S6, Lai et al., Anal. Bioanal. Chem. 377, 208–213, 2003; S7, Sanbe et al., J. Pharm. Biomed. Anal. 30, 1835–1844, 2002; S8, Haginaka et al., J. Pharm. Biomed. Anal. 22, 899–907, 2000; S9, Haginaka et al., Anal. Sci. 19, 39–42, 2003; S10, Haginaka et al., Anal. Chem. 72, 5206–5210, 2000; S11, Haginaka et al., J. Chromatogr. B 804, 19–24, 2004

number of variables (in our case molecular descriptors and imprinting factor, respectively) presented in the dataset. On the other hand, the numbers of nodes to use for the hidden layer are obtained through trial-and-error. A schematic representation of the layers of neural network is illustrated in Scheme 1.

Each node of the hidden and output layer contains two components, namely the summation function and the transfer function. Summation function is computed from the weighted sum of all input node entering each hidden node and it consolidates the weights of various input to the neuron into a single value that can be passed on to the transfer function for further processing. The role of the transfer function is to translate the summed information into outputs. One of the most popular transfer functions and the one used in this work is the sigmoid transfer function.

The connections among nodes of the various layers are assigned numerical values known as weights, which express the relative strength of the input data. The learning process starts with a random seeding of the connection weights [27, 28] and signals are propagated from the input layer through the hidden layer to the output layer. A neural network is trained by adjusting the weights until they reach an optimal set where the predicted output is as close as possible to the actual output for as many input samples presented in the dataset. There are many learning algorithms in neural network and each method differ in the way weights are adjusted. In this study, we used the back-propagation algorithm and it involves the correction of weights starting from the output layer and working its way backwards towards the input layer. Back-propagation is a supervised learning method meaning that it requires both the input and output to be known beforehand. Afterwards the weights are adjusted accordingly with respect to the error, which are calculated from the difference between the actual and predicted value [29].

Neural network

Artificial neural network calculations were performed with the back-propagation implementation of WEKA 3.4.3 [30] on a personal computer running Windows XP with Intel Pentium 4 3.0 GHz CPU and 1024 MB of RAM. The neural

Table 2. The irregularly-sized MIP dataset obtained from the literature.

No.	Template	Functional monomer	MIP k'	NIP k'	IF	Mobile phase composition	Source
1	Picolinamide	MAA	1.71	1.67	1.02	Acetonitrile	S1
2	Isonicotinamide	MAA	11.7	2.49	4.70	Acetonitrile	S1
3	Nicotinamide	MAA	13.1	2.65	4.94	Acetonitrile	S1
4	p-hydroxybenzoic acid	AAm	2.75	0.66	4.17	Acetonitrile/ Acetic acid (99.75/0.25, v/v)	S1
5	p-hydroxyphenylacetic acid	AAm	1.85	0.41	4.51	Acetonitrile/ Acetic acid (99.75/0.25, v/v)	S1
6	p-hydroxyphenylpropionic acid	AAm	1.02	0.2	5.10	Acetonitrile/ Acetic acid (99.75/0.25, v/v)	S1
7	Quercetin	AAm	48.6	2.6	18.70	Methanol	S2
8	Histamine	MAA	5.64	1.29	4.37	Acetonitrile/ Acetic acid (80/20, v/v)	S3
9	17β-Estradiol	MAA	1.59	0.43	3.70	Acetonitrile/ Acetic acid (99/1, v/v)	S4
10	Dienestrol	MAA	0.22	0.13	1.69	Acetonitrile/ Acetic acid (99/1, v/v)	S4
11	Gentisic acid	AAm	8.23	4.13	1.99	Acetonitrile/ Acetic acid (99.5/0.5, v/v)	S5
12	2,4,5-trichlorophenoxyacetic-acid	4-VP	4.39	0.44	9.98	Methanol/ Acetic acid (99/1, v/v)	S6
13	Chloramphenicol	2-DAEMA	12.83	2.59	4.95	Acetonitrile	S7
14	Chloramphenicol	2-VP	13.21	0.6	22.02	Acetonitrile	S7
15	Chloramphenicol	AA	4.22	3.06	1.38	Acetonitrile	S7
16	Chloramphenicol	MAA	0.33	0.29	1.14	Acetonitrile	S7
17	Chloramphenicol	2-HEMA	0.27	0.16	1.69	Acetonitrile	S7
18	Cholesterol	2-MAOEP	15.8	6.4	2.47	n-hexane	S8
19	β-estradiol	2-MAOEP	15	11	1.36	n-hexane	S8
20	Sulfamethazine	MAA	2.4	1.25	1.92	Acetonitrile	S9
21	Sulfamethoxazole	MAA	1.08	1	1.08	Acetonitrile	S9
22	Sulfamethazine	AAm	1.25	1.37	0.91	Acetonitrile	S9
23	Sulfamethoxazole	AAm	1.47	1.26	1.17	Acetonitrile	S9
24	Sulfamethazine	4-VP	0.36	0.24	1.50	Acetonitrile	S9
25	Sulfamethoxazole	4-VP	1.5	0.37	4.05	Acetonitrile	S9
26	Cortisol	MAA	4.37	0.45	9.71	Chloroform/ Acetic acid (99.5/0.5, v/v)	S10
27	Indoleacetic acid	MAA	13	8.4	1.55	Chloroform/ Acetic acid (99.9/0.1, v/v)	S11
28	Indoleacetic acid	N,N-DMAEMA	50	34	1.47	Chloroform/ Acetic acid (99.9/0.1, v/v)	S11
29	Alloxan	2,6-BAAP	4.88	2.63	1.86	Chloroform/ Methanol (95/5, v/v)	S12
30	Cyclobarbital	2,6-BAAP	12.4	0.34	36.47	Chloroform	S12
31	Hexobarbital	2,6-BAAP	2.46	0.34	7.24	Chloroform	S12
32	BOC-D-Ala-L-Ala-pNA	L-Val	2.09	1.06	1.97	Hexane/ Chloroform (60/40, v/v)	S13
33	BOC-D-Phe-L-Ala-pNA	L-Val	2.95	0.54	5.46	Hexane/ Chloroform (60/40, v/v)	S13
34	Nicotine	MAA	N/A	N/A	11.00	Acetonitrile/ Acetic acid (95/5, v/v)	S14
35	Nicotine	2-TFMAA	N/A	N/A	7.50	Acetonitrile/ Acetic acid (95/5, v/v)	S14
36	2-nitrophenol	4-VP	0.04	0.05	0.80	Acetonitrile	S15
37	3-nitrophenol	4-VP	0.6	0.41	1.46	Acetonitrile	S15

38	4-nitrophenol	4-VP	1.9	0.65	2.92	Acetonitrile	S15
39	Cinchonidine	DBOMAT	0.92	0.32	2.88	Chloroform/ Acetic acid (99.9/0.1, v/v)	S16
40	Cinchonine	DBOMAT	1.56	0.24	6.50	Chloroform/ Acetic acid (99.9/0.1, v/v)	S16
41	E-Piceatannol	4-VP	9.71	2.05	4.74	Methanol	S17
42	Quercetin	4-VP	6.98	2.07	3.37	Methanol	S17
43	Terbutylazine	MAA	18	1.2	15.00	Acetonitrile	S18
44	p-hydroxybenzoic acid	4-VP	45	3.15	14.29	Acetonitrile/ Acetic acid (99.75/0.25, v/v)	S19
45	p-hydroxyphenylacetic acid	4-VP	7.02	1.9	3.69	Acetonitrile/ Acetic acid (99.75/0.25, v/v)	S19
46	p-hydroxyphenylpropionic acid	4-VP	4.91	1.71	2.87	Acetonitrile/ Acetic acid (99.75/0.25, v/v)	S19
47	Salicylaldehyde	4-VP	2.28	0.51	4.47	Methanol	S20
48	Testosterone	MAA	6.03	0.53	11.38	Acetonitrile	S21
49	Testosterone	2-DAEMA	0.46	0.46	1.00	Acetonitrile	S22
50	β-estradiol	2-DAEMA	2.78	2.59	1.07	Acetonitrile	S22
51	Testosterone-propionate	2-DAEMA	0.33	0.33	1.00	Acetonitrile	S22
52	Progesterone	2-DAEMA	0.27	0.27	1.00	Acetonitrile	S22
53	Estrone	2-DAEMA	1.92	1.72	1.12	Acetonitrile	S22
54	Testosterone	MAA	4.51	1.26	3.58	Acetonitrile	S22
55	β-estradiol	MAA	1.52	1.19	1.28	Acetonitrile	S22
56	Testosterone-propionate	MAA	0.59	0.46	1.28	Acetonitrile	S22
57	Progesterone	MAA	0.79	0.59	1.34	Acetonitrile	S22
58	Estrone	MAA	0.53	0.46	1.15	Acetonitrile	S22
59	4-chlorophenol	4-VP	1.01	0.71	1.42	Acetonitrile/ Acetic acid (99.7/0.3, v/v)	S23
60	Bisphenol A	4-VP	6.17	0.8	7.71	Acetonitrile/ Methanol (85/15, v/v)	S24
61	Benzo[a]pyrene	4-VP	11.59	1.36	8.52	Acetonitrile/ Dichloromethane (85/15, v/v)	S25
62	Phenytoin	MAAm	7.16	1.3	5.51	Acetonitrile	S26
63	Bentazone	MAA	0.84	0.8	1.05	Acetonitrile	S27
64	Bentazone	4-VP	11.8	8.09	1.46	Acetonitrile	S27
65	Ephedrine	MAA	9.46	3.3	2.87	Chloroform/ Acetic acid (99.7/0.3, v/v)	S28
66	Ephedrine	2-HEMA	0.8	0.3	2.67	Chloroform/ Acetic acid (99.7/0.3, v/v)	S28
67	Ephedrine	AAm	2.59	1.4	1.85	Chloroform/ Acetic acid (99.7/0.3, v/v)	S28
68	Ephedrine	2-VP	0.1	0.1	1.00	Chloroform/ Acetic acid (99.7/0.3, v/v)	S28
69	11α-hydroxyprogesterone	MAA	6.15	0.31	19.84	Dichloromethane/ Acetic acid (99.9/0.1, v/v)	S29
70	Corticosterone	MAA	6.6	0.32	20.63	Dichloromethane/ Acetic acid (99.5/0.5, v/v)	S29
71	Hexestrol	2-DAEMA	2.797	0.438	6.39	Acetonitrile	S30
72	5-fluorouracil	2,6-BAP	6.9	0.3	23.00	Acetonitrile	S31
73	1-ethoxymethyl-5-fluorouracil	2,6-BAAP	2.2	0.3	7.30	Acetonitrile	S31

Table 2. Continued

No.	Template	Functional monomer	MIP k'	NIP k'	IF	Mobile phase composition	Source
74	Uracil	2,6-BAAP	1.2	0.3	4.00	Acetonitrile	S31
75	Bupivacaine	MAA	0.682	0.0365	18.68	Chloroform/ Acetic acid (95/5, v/v)	S32

MIP k', Retention factor of molecularly imprinted polymer; NIP k', Retention factor of non-imprinted polymer; IF, Imprinting factor; N/A, not available 2-HEMA, 2-hydroxyethyl-methacrylate; 2-MAOEP, 2-(methacryloyloxy)ethyl-phosphate; 2-VP, 2-vinylpyridine; 2-TFMAA, 2-(Trifluoromethyl)acrylic acid; 4-VP, 4-vinylpyridine; 2-DAEMA, 2-(diethylamino)ethyl methacrylate; 2,6-BAAP, 2,6-bis(acrylamido)pyridine; AAm, Acrylamide; AA, Allylamine; DBOMAT, dibenzyl (2R,3R)-O-monoacryloyl tartate; L-Val, L-Valine derivative; MAA, methacrylic acid; MAAm, methacrylamide; N,N-DMAEMA, N,N-dimethylaminoethyl methacrylate

S1, Wu et. al., Analyst 128, 944–949, 2003; S2, Xie et. al., J. Chromatogr. A 934, 1–11, 2001; S3, Allender et. al., Int. J. Pharm. 195, 39–43, 2000; S4, Ye et. al., Analyst 126, 760–765, 2001; S5, Zhang et. al., Anal. Chim. Acta 450, 53–61, 2001; S6, Baggiani et. al., J. Chromatogr. A 938, 35–44, 2001; S7, Suarez-Rodriguez et. al., Biosen. Bioelectron. 16, 955–961, 2001; S8, Kugimiya et. al., J. Chromatogr. A 938, 131–135, 2001; S9, Zheng et. al., Microchem. J. 69, 153–158, 2001; S10, Baggiani et. al., Talanta 51, 71–75, 2000; S11, Kugimiya et. al., Anal. Chim. Acta 395, 251–255, 1999; S12, Yano et. al., Anal. Chim. Acta 363, 111–117, 1998; S13, Anal. Chim. Acta 357, 91–98, 1997; S14, Matsui et. al., Anal. Chim. Acta 343, 1–4, 1997; S15, Huang et. al., J. Mol. Recognit. 16, 406–411, 2003; S16, Knutsson et. al., J. Mol. Recognit. 11, 87–90, 1998; S17, Zhu et. al., Anal. Chem. 75, 6381–6387, 2003; S18, Ferrer et. al., Anal. Chem. 72, 3934–3941, 2000; S19, Sun et. al., J. Mol. Recognit. 14, 388–392, 2001; S20, Yang et. al., J. Mol. Recognit. 18, 103–108, 2005; S21, Rachkov et. al., Polym. Adv. Technol. 9, 511–519, 1998; S22, Cheong et. al., J. Polym. Sci. Pol. Chem. 36, 1725–1732, 1998; S23, Caro et. al., J. Chromatogr. A 995, 233–238, 2003; S24, Vicente et. al., Anal. Bioanal. Chem. 380, 115–122, 2004; S25, Lai et. al., Anal. Chim. Acta 522, 137–144, 2004; S26, Bereczki et. al., J. Chromatogr. A 930, 31–38, 2001; S27, Baggiani et. al., Anal. Commun. 36, 263–266, 1999; S28, Piletsky et. al., Analyst 126, 1826–1830, 2001; S29, Ramstrom et. al., Anal. Commun. 35, 9–11, 1998; S30, Tarbin et. al., Anal. Commun. 36, 105–107, 1999; S31, Kugimiya et. al., Analyst 126, 772–774, 2001; S32, Karlsson et. al., Analyst 129, 456–462, 2004

network used in this study consists of three layers: 1) input layer, 2) hidden layer, and 3) output layer. The network architecture of the uniformly sized MIP dataset is $35\times Y\times 1$, which means 35 neurons (nodes) in the input layer, Y neurons in the hidden layer, and 1 neuron in the output layer. The network architecture of the irregularly sized MIP data set is $5\times Y\times 1$. The input layer consists of nodes that represent the mobile phase descriptors and TAE molecular descriptors of both the template and functional monomer. Therefore, for the uniformly sized MIP dataset there are 16 template TAE molecular descriptors +16 functional monomer TAE molecular descriptors +3 mobile phase descriptors = 35 descriptors or nodes. As for the irregularly sized MIP dataset there are 2 template TAE molecular descriptors +2 functional monomer TAE molecular descriptors +1 mobile phase descriptor = 5 descriptors or nodes. The node in the output layer represents the imprinting factor.

As a rule of thumb, optimization of the neural network condition was performed by trial-and-error by adjusting various parameters. These included the number of nodes in the hidden layer, the learning epoch size, and the learning rate and momentum constants. The learning rate (η) constant determines the speed at which the weights change, while the momentum (μ) constant prevents sudden changes in attaining the solution [29]. One complete cycle of data propagated in a feed-forward manner through the layers of the neural network is referred to as an epoch.

Root mean square error (RMS) was used as a measure of the prediction error by the trained model and is calculated with the following equation:

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}} \tag{5}$$

where p_i is the predicted output, a_i is the actual output, and n is the number of compounds in the dataset. The RMS of various parameters was calculated and the optimal value to use for each parameter was those possessing low RMS value. Firstly, the optimal number of TAE molecular descriptors was selected by plotting RMS as a function of the number of TAE molecular descriptors. Then, the optimal number of nodes in the hidden layer was determined by varying the

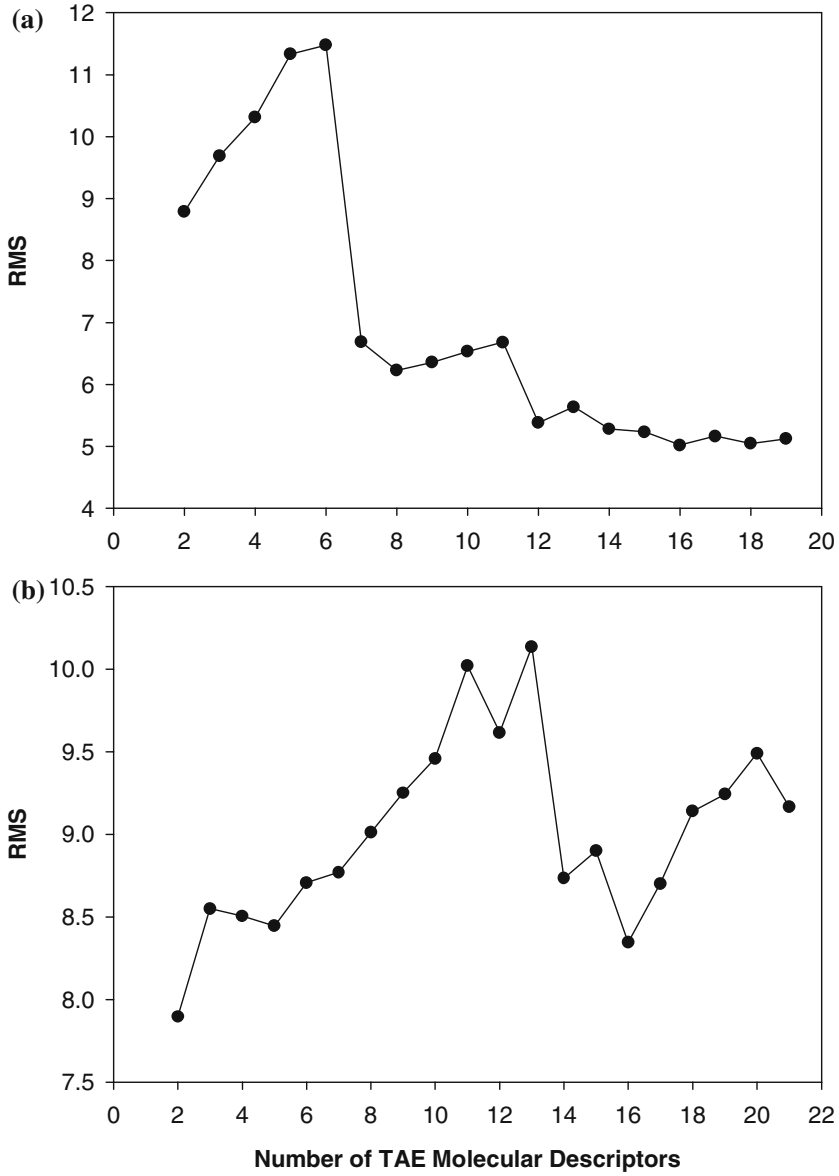


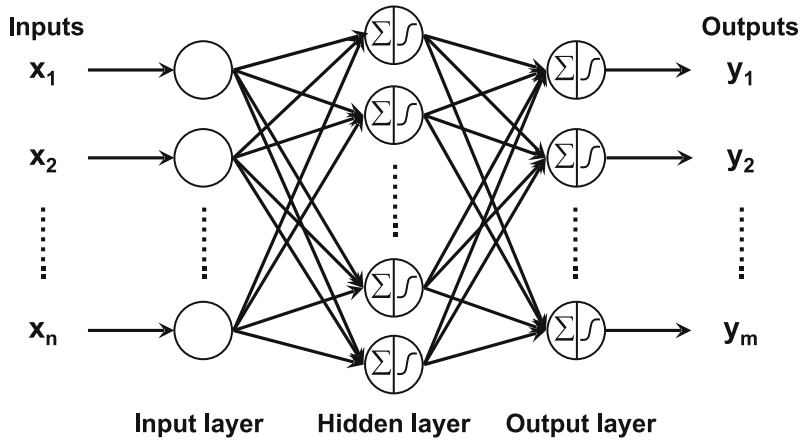
Figure 1. RMS as a function of the number of TAE molecular descriptors for the uniformly-sized MIP dataset (a) and the irregularly-sized MIP dataset (b).

number of hidden nodes and observing the RMS. The number of nodes that gave the lowest RMS was chosen. Calculations were performed with leave-one-out cross-validation and the average RMS of 10 runs was used as the measure of predictive error. Next, the best training time was determined by plotting a curve of the RMS versus the training time at the optimal number of hidden nodes. As before, the training time that gave low RMS was chosen. Finally, optimal learning rate and momentum were determined by varying both terms and the pair giving low RMS was chosen.

Results and discussion

Network optimization

The optimal architecture of the neural network was obtained by adjusting various parameters by trial-and-error. The same practice is customary to



Scheme 1. Schematic representation of a three-layer feed-forward back-propagation neural network used in this study. Circles represent nodes (neurons), while the connection between nodes represents weights. The summation and sigmoid symbol found inside the nodes of the hidden and output layer represents summation and sigmoid transfer function, respectively.

the field of neural network since there is no rule that justifies the optimal network architecture to use. The empirically determined parameters include: the number of nodes in the hidden layer, when to stop training, and the learning rate and momentum. Selection of optimal parameters relied on the use of RMS as the error function, where parameters that yield low RMS were chosen. For each of the parameter calculations, 10 runs were performed and the averages of the RMS were calculated.

Prior to the network optimization, both datasets were subjected to descriptor reduction by UFS by varying the user-defined parameter, $R^2_{\max}$, between 0 and 1 to obtain a reduced set of TAE molecular descriptor in the range of 2 to 19 for the uniformly sized MIP dataset and 2 to 21 for the irregularly sized MIP dataset. After the TAE molecular descriptors were reduced by UFS, the value with the lowest RMS was chosen as optimal. Thus, 16 TAE molecular descriptors were used for the uniformly sized MIP dataset while 2 TAE molecular descriptors were used for the irregularly sized MIP dataset.

In order to optimize the size of the learning epoch as well as to avoid overtraining, early stopping was used in our network training. Early stopping refers to the practice of stopping learning after a certain predefined number of errors are reached on the validation set during network training [31, 32]. Leave-one-out cross-validation refers to the removal of one sample of the dataset for use as the test set while using the rest as training set. This is repeated until all samples of the dataset are used as the test set. In our work, one sample of the dataset was left out as the test set, while the remaining dataset were randomly split into two portions: 90% as the training set and 10% as the validation set. Training sets were used by the neural network software to create a predictive model that contains the essential information (in our case the appropriate weights of the network connection) necessary to predict the output. The sample that was left out was then used as the test set in order to assess the accuracy of the predictive model. On the other hand, when the samples of the dataset were used as both the training and test set, the predictive model is merely used to *recall* the samples of the dataset. By comparing both the performance of the test set and training set, the optimal time to stop training can be extrapolated. When the performance on the training set becomes better while the performance on the test set starts to deteriorate, such phenomenon is called *overtraining* and training is usually stopped before this occurs. Due to the limited data, early stopping was employed whereby the validation set were used to monitor the training of the neural network and training were stopped when 20 errors on the validation set were reached.

The optimal number of nodes in the hidden layer was determined by varying the number of nodes from 1 to 25. As can be seen from the plot of RMS versus Hidden nodes (Figure 2), the optimal

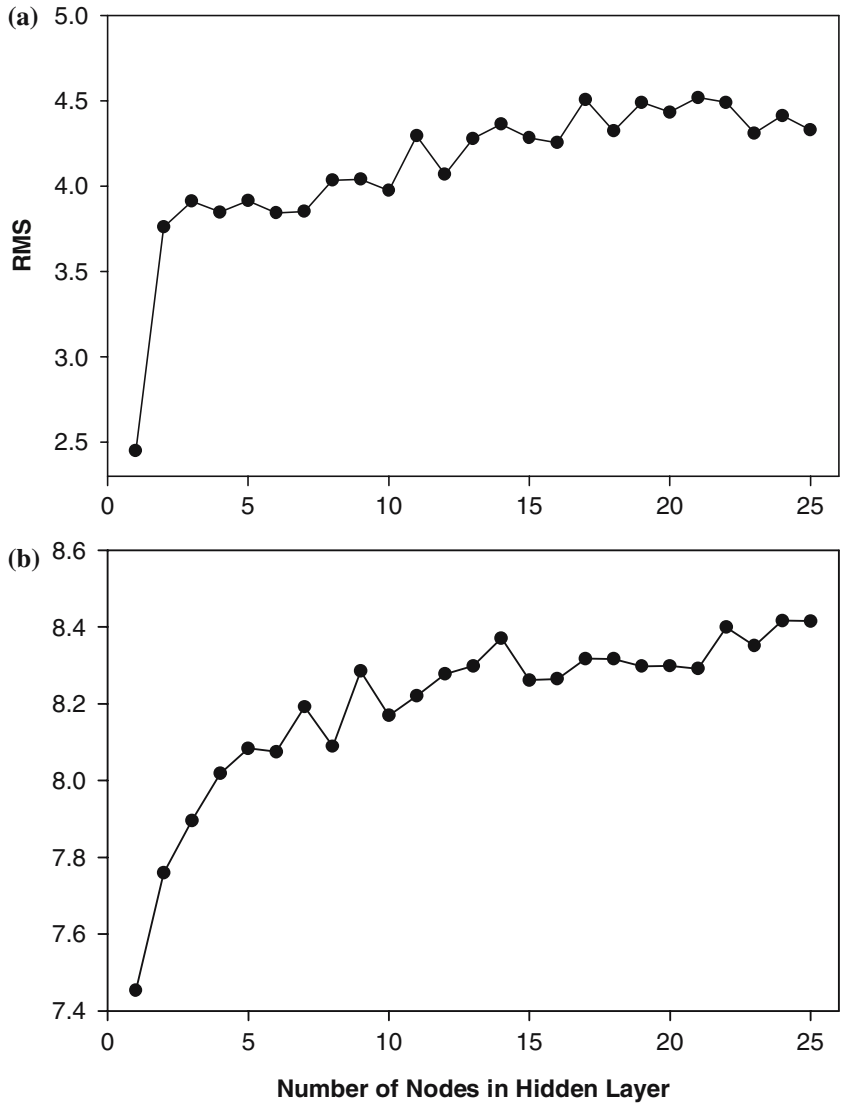


Figure 2. RMS as a function of the number of nodes in the hidden layer for the uniformly-sized MIP dataset (a) and the irregularly-sized MIP dataset (b).

number of nodes in the hidden layer is 1 for both the uniformly sized MIP dataset and irregularly sized MIP dataset.

The best learning time or the optimal epoch size were determined by plotting a graph of the RMS as a function of the learning epoch size. The first RMS was recorded after the first epoch and subsequent RMS were measured after every 50 epochs. The optimal learning epoch was found to be 500 for the uniformly sized MIP dataset and 300 for the irregularly sized MIP dataset (Figure 3). The learning epoch for the uniformly sized MIP dataset was left at the default 500 epochs since no increase in RMS was seen beyond 500 epochs, while 300 epochs was chosen for the irregularly sized MIP dataset because the RMS increased slightly beyond 300 epochs, signifying overtraining.

The optimal learning rate and momentum were selected by making a contour plot of the RMS as a function of the learning rate and momentum (Figure 4). Each line found in the contour plot represents a constant value of the RMS error while the shaded boxes corresponds to the RMS values

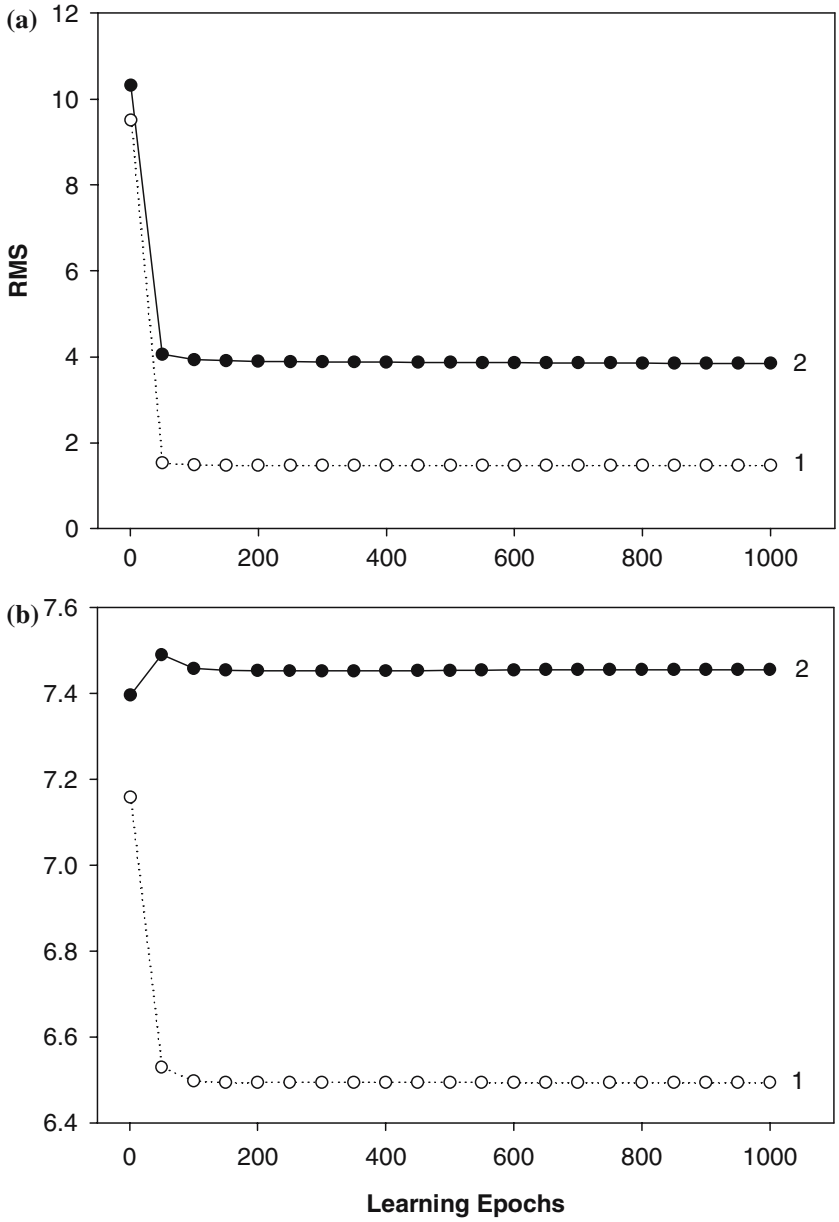


Figure 3. RMS as a function of the number of learning epochs for the uniformly-sized MIP dataset (a) and the irregularly-sized MIP dataset (b). Curves 1 and 2 represent the training set and test set, respectively.

obtained from the learning procedures and fitted onto the same surface model [29]. As we can see from the contour plot, the good learning rate and momentum lies in the middle left region of the plot for the uniformly sized MIP dataset (Figure 4a) and lower left region for the irregularly sized MIP dataset (Figure 4b). Thus, the optimal learning rate and momentum is 0.2 and 0.5, respectively, for the uniformly sized MIP dataset, and 0.1 and 0.1, respectively, for the irregularly sized MIP dataset.

Prediction of imprinting factor using artificial neural networks

Through the aforementioned optimizations, the optimal hidden nodes, learning epoch size, learning rate and momentum were found to be 1, 500, 0.2, and 0.5 for the uniformly sized MIP dataset and 1, 300, 0.1, and 0.1 for the irregularly sized MIP dataset. The prediction results are shown in Figure 5 as a plot of the data points as a function

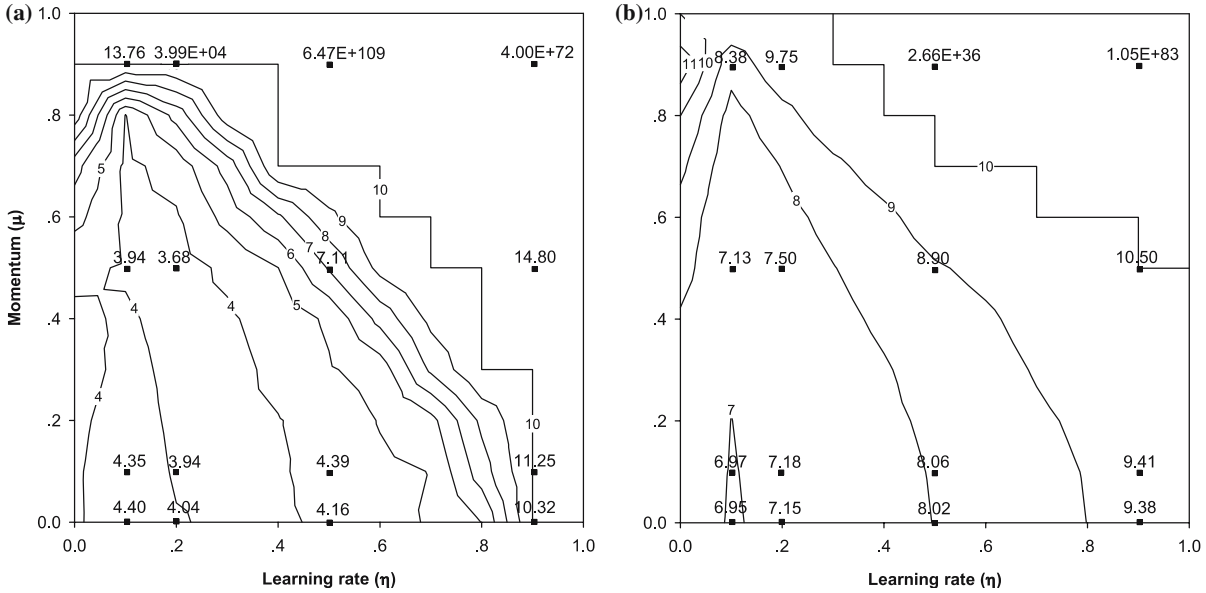


Figure 4. Contour plot of RMS in test set versus learning rate (η) and momentum (μ) for the uniformly-sized MIP dataset (a) and the irregularly-sized MIP dataset (b). Each line represents constant value of the RMS, while shaded boxes represent RMS values obtained from the training procedure and fitted onto the same surface model of the contour plot.

of the calculated and literature imprinting factor. It can be seen that the uniformly sized MIP dataset gave reasonably good predictive power with a correlation coefficient of $r = 0.946$ and RMS of 2.944. On the other hand, the irregularly sized MIP dataset yielded rather poor predictive power with a correlation coefficient of $r = 0.382$ and RMS of 6.123. Both of the datasets were obtained through leave-one-out cross-validation and the RMS was calculated by averaging values obtained from 10 runs.

In the molecular imprinting literature, uniformly sized MIP particles are preferred over irregularly sized MIP particles due to 'large surface area, monodispersity, colloidal stability' of the uniformly sized MIP particles. In addition, the uniformly sized MIP particles facilitate better template removal and allow fast binding kinetics due to their surface exposed binding sites [33]. These differences could be attributed to the fact that the polymers were prepared by different methods, especially those created by bulk polymerization are known to lack control over the polymerization process [16]. This lack of control would probably lead to the observed differences in the size and shape of the MIP particles as well as heterogeneity in the binding sites. Non-covalent molecular imprinting are generally expected to possess heterogeneous binding sites, but great improvements in the polymerization process have made it possible to generate non-covalent imprinted polymers containing near homogeneous binding sites. Cacho and co-workers prepared uniformly sized MIP particles by precipitation polymerization with experimentally determined heterogeneity index (m) close to 1 for the propazine-imprinted polymer [34] and 1 for the fenuron-imprinted polymer [35], where 1 represents a homogeneous material and heterogeneous material have values between 0 and 1. In addition to the near homogeneous binding site distribution, there are also fast diffusion mass transfer kinetics and higher association constants [34–36]. Given that the uniformly sized MIP particles generated by the aqueous suspension and multi-step swelling polymerization method contain predominantly surface exposed binding sites, the association constant of their binding sites would be expected to be of comparable magnitude due to the ease of accessibility and low heterogeneity of the binding sites. In addition, since the total number of binding sites for the uniformly sized MIP particles is less than that of the irregularly sized MIP particles but the capacity factor for the uniformly sized MIP particles are many magnitude greater, it follows that the binding sites possess higher

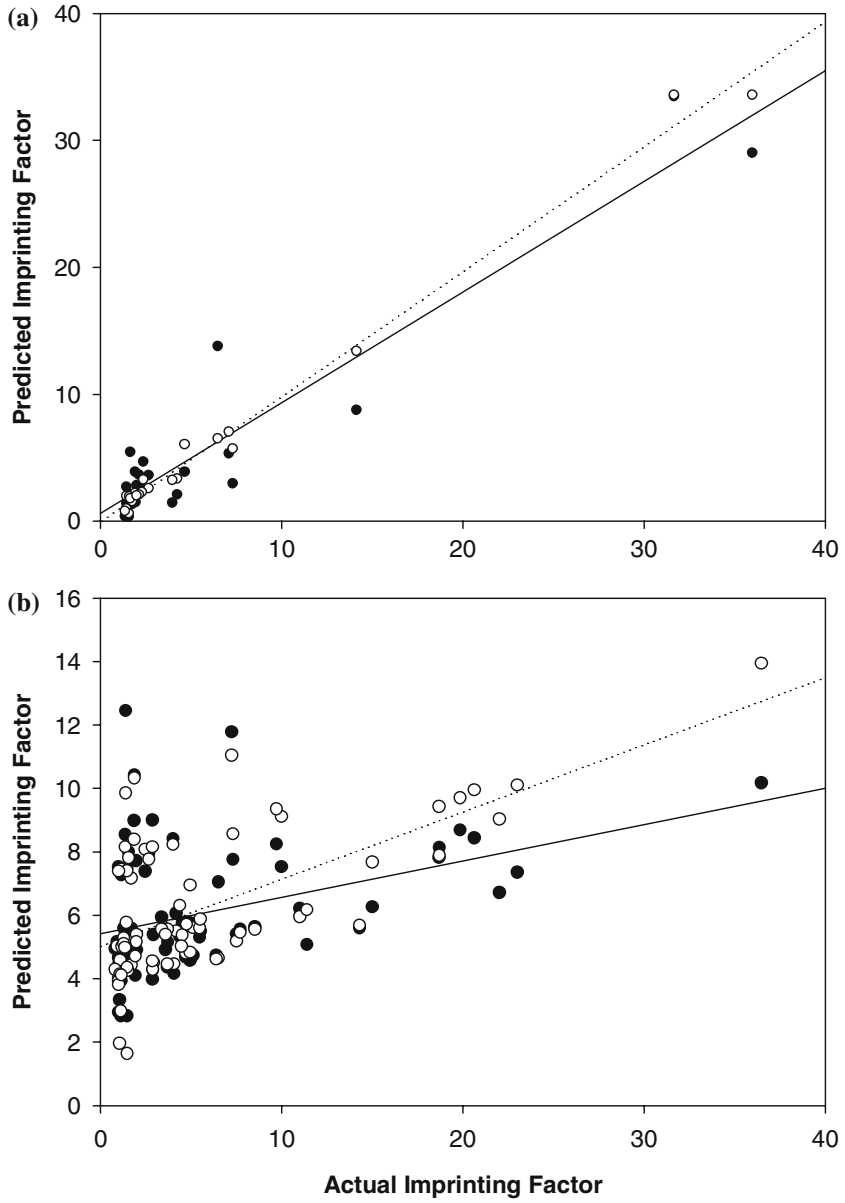


Figure 5. Plot of the predicted vs. actual imprinting factor for the training set (○; regression line is represented as a dotted line) and leave-one-out cross-validated test set (●; regression line is represented as a solid line) of the uniformly-sized MIP dataset (a) and the irregularly-sized MIP dataset (b). The uniformly-sized MIP dataset had correlation coefficient of $r = 0.946$ and RMS of 2.944 for the leave-one-out cross-validated test set and $r = 0.995$ and RMS of 0.896 for the training set. The irregularly-sized MIP dataset had $r = 0.382$ and RMS of 6.123 on the leave-one-out cross-validated test set and $r = 0.628$ and RMS of 5.500 for the training set.

association constants [34]. Besides the surface exposed binding sites, the irregularly sized MIP particles also contain binding sites that are buried deep within the macromolecular polymer matrix of the particle and the accessibility of template molecules to these binding sites are rather difficult. The control over the particle size of uniformly sized MIP particles allows the production of monodisperse population of MIP particles thereby generating particles with comparable number of binding sites [16, 33]. On the other hand, uniformly sized MIP particles are of variable dimensions, therefore producing polymers with variable number of binding sites. Moreover, the

monodisperse uniformly sized MIP particles pack efficiently in chromatographic columns; give good flow properties, low back pressure, and good column efficiencies and peak shapes [36–41].

Apart from the differences in the shape and size of the MIP particles, it is plausible to assume that the different polymerization methods also give rise to variation in the number and association constants of the binding sites. In the bulk polymerization method, the grinding step essentially helps break the bulky monolithic polymers into smaller particles. The grinding step also causes the destruction of binding sites, which in effect reduces the number of binding sites, and the creation of fragmented binding sites [37] leads to erratic association constant, which can be attributed to the 'wide range of affinities and selectivities' of the binding sites [42]. Furthermore, the extent of the binding site destruction and fragmentation is unpredictable and is likely to reduce the predictive power. The polymerization methods used to produce uniformly sized MIP particles yield reproducible results [37] while methods used to generate irregularly sized MIP particles generate particles of unpredictable nature.

The influence of different polymerization method on the predictive model is clearly seen in the uniformly sized MIP dataset upon omission of data prepared by aqueous micro-suspension polymerization. Such removal did in fact increase the correlation coefficient of the dataset of uniformly sized MIP prepared solely by the multi-step swelling and polymerization from 0.946 to 0.954 and reduce the RMS from 2.944 to 2.932 (data not shown). The improvements could be attributed to the differences that are found inherently in the two polymerization methods. Since the two methods produce particles of high-quality, the improvements seen were negligible, whereas the differences observed in the predictive power of the uniformly sized MIP and irregularly sized MIP indicated that both polymers differed greatly in terms of their morphology and binding site properties, which have great influence on the predictive model.

Conclusion

In this study, we have explored the application of artificial neural network as a novel approach in the field of molecular imprinting, particularly in the prediction of imprinting factor of molecularly imprinted polymers. The strength of interaction between any given pair of template and functional monomer can be inferred from the magnitude of the imprinting factor. The uniformly sized MIP dataset exhibits better predictive power than that of the irregularly sized MIP dataset, which can be attributed to the more consistent and uniform nature of the uniformly sized MIP particles. Further work is under way to explore the application of artificial neural networks and related data mining techniques in other facets of molecular imprinting.

Acknowledgements

C.N. is grateful for the Royal Golden Jubilee Ph.D. Scholarship under V.P. supervision from The Thailand Research Fund. This project was also partially supported by the Thailand Toray Science Foundation (TTSF) and a grant from the annual budget of Mahidol University (B.E.2548).

References

1. Ye, L. and Mosbach, K., J. Incl. Phenom. Macrocycl. Chem., 41 (2001) 107.
2. Dobashi, A., Nishida, S., Kurata, K. and Hamada, M., Anal. Sci., 18 (2002) 35.
3. Lin, T.Y., Hu, C.H. and Chou, T.C., Biosens. Bioelectron., 20 (2004) 75.
4. Piacham, T., Josell, Å., Arwin, H., Prachayasittikul, V. and Ye, L., Anal. Chim. Acta, 536 (2005) 191.
5. Vlatakis, G., Anderson, L.I., Muller, R. and Mosbach, K., Nature, 361 (1993) 645.
6. Piacham, T., Isarankura Na Ayudhya, C., Prachayasittikul, V., Bülow, L. and Ye, L., Chem. Commun., (2003) 1254.
7. Mosbach, K., Yu, Y., Andersch, J. and Ye, L., J. Am. Chem. Soc., 123 (2001) 12420.
8. Sellergren, B., Lepistoe, M. and Mosbach, K., Am. Chem. Soc., 110 (1988) 5853.
9. Duffy, D.J., Das, K., Hsu, S.L., Penelle, J., Rotello, V.M. and Stidham, H.D., Polym. Mat. Sci. Eng., 82 (2000) 69.
10. Nicholls, I.A., Adbo, K., Andersson, H.S., Andersson, P.O., Ankarloo, J., Hedin-Dahlstrom, J., Jokela, P., Karlsson, J.G., Olofsson, L. and Rosengren, J., Anal. Chim. Acta, 435 (2001) 9.
11. Chianella, I., Lotierzo, M., Piletsky, S.A., Tothill, I.E., Chen, B., Karim, K. and Turner, A.P., Anal. Chem., 74 (2002) 1288.
12. Subrahmanyam, S., Piletsky, S.A., Piletska, E.V., Chen, B., Karim, K. and Turner, A.P., Biosens. Bioelectron., 16 (2001) 631.

13. Piletsky, S.A., Karim, K., Piletska, E.V., Day, C.J., Freebairn, K.W., Legge, C. and Turner, A.P., Analyst, 126 (2001) 1826.
14. Wu, L. and Li, Y., J. Mol. Recognit., 17 (2004) 567.
15. Wu, L., Sun, B., Li, Y. and Chang, W., Analyst, 128 (2003) 944.
16. Pérez-Moral, N. and Mayes, A.G., Anal. Chim. Acta, 504 (2004) 15.
17. MARVIN, Version 3.5.4, ChemAxon Ltd., Budapest, Hungary, http://www.chemaxon.com/marvin.
18. RECON, Version 5.5, Rensselaer Polytechnic Institute, Troy, New York, USA, http://www.chem.rpi.edu/chemweb/recondoc.
19. Breneman, C.M. and Rhem, M., Comput. Chem., 18 (1997) 182.
20. Breneman, C.M., Thompson, T.R., Rhem, M. and Dung, M., Comput. Chem., 19 (1995) 161.
21. Lide, D.R. CRC Handbook of Chemistry and Physics: A Ready-Reference Book of Chemical and Physical Data, 71th ed.,. CRC Press, Inc, Florida, 1990.
22. Dumanovic, D., Kosanovic, D.J., Arkakovic, D. and Jovanovic, J., Pharmazie, 47 (1992) 603.
23. Chien, Y.W., J. Parenter. Sci. Technol., 38 (1984) 32.
24. Prakongpan, S. and Nagai, T., Chem. Pharm. Bull., 32 (1984) 340.
25. McNaught, A.D. and Wilkinson, A., IUPAC Compendium of Chemical Terminology, 2nd ed., http://www.iupac.org/goldbook/I03180.pdf.
26. Whitley, D.C., Ford, M.G. and Livingstone, D.J., J. Chem. Inf. Comput. Sci., 40 (2000) 1160.
27. Niculescu, S.P., J. Mol. Struct., 622 (2003) 71.
28. Agatonovic-Kustrin, S., Zecevic, M. and Zivanovic, L., J. Pharm. Biomed. Anal., 21 (1999) 95.
29. Zupan, J. and Gasteiger, J. Neural Networks in Chemistry and Drug Design, 2nd ed.,. Wiley-VCH, Weinheim, 1999.
30. Witten, I.H. and Frank, E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers, San Francisco, 2000.
31. Zhang, R., Yan, A., Liu, M., Liu, H. and Hu, Z., Chemometr. Intell. Lab. Syst., 45 (1999) 113.
32. Loukas, Y.L., J. Chromatogr. A, 904 (2000) 119.
33. Pérez-Moral, N. and Mayes, A.G., Bioseparation, 10 (2001) 287.
34. Cacho, C., Turiel, E., Martin-Esteban, A., Pérez-Conde, C. and Cámara, C., J. Chromatogr. B, 802 (2004) 347.
35. Tamayo, F.G., Casillas, J.L. and Martin-Esteban, A., Analyst, 128 (2003) 137.
36. Baggiani, C., Baravalle, P., Anfossi, L. and Tozzi, C., Anal. Chim. Acta, 542 (2005) 125.
37. Mayes A.G., 2001. Polymerisation techniques for the formation of imprinted beads In: Sellergren B., (Eds) Molecularly Imprinted Polymers: Man-made Mimics of Antibodies and their Applications in Analytical Chemistry (Chapter 12) Elsevier, Amsterdam pp. 305–324.
38. Fu, Q., Sanbe, H., Kagawa, C., Kunimoto, K.K. and Hagina, , J. Anal. Chem., 75 (2003) 191.
39. Lai, J.P., Cao, X.F., Wang, X.L. and He, X.W., Anal. Bioanal. Chem., 372 (2002) 391.
40. Haginaka, J. and Kagawa, C., Anal. Sci., 19 (2003) 39.
41. Haginaka, J. and Kagawa, C., J. Chromatogr. A, 948 (2002) 77.
42. Umpleby, R.J. II, Baxter, S.C., Rampey, A.M., Rushton, G.T., Chen, Y. and Shimizu, K.D., J. Chromatogr. B, 804 (2004) 141.