Research Article

Principal Component Analysis Coupled with Artificial Neural Networks for Therapeutic Indication Prediction of Thai Herbal Formulae

Lawan Sratthaphut[1*], Samart Jamrus[1], Suthikarn Woothianusorn[1] and Onoomar Toyama[2]

[1]Department of Health-Related Informatics, [2]Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Silpakorn University, Nakhon Pathom, Thailand
[*]Corresponding author E-mail address: lawan.s@su.ac.th

Received October 8, 2012; Accepted November 26, 2012

Abstract

This study illustrated the principal component analysis coupled with artificial neural networks (PC-ANN) as a useful tool in therapeutic indication prediction of Thai herbal formulae official in the National List of Essential Medicine 2011 and the National Traditional Household Remedies. A set of 71 herbal formulae from the National List of Essential Medicine 2011 and the National Traditional Household Remedies and 19 formulae without therapeutic indication was used as a training set, a monitoring set and a validation set. The performance of the model was measured by the percentage of "correctly classified", True Positive rate and False Positive rate of the PC-ANN model. The results suggested that principal component analysis technique could condense all of the variables in which there were interrelated, into a few principal components, while retaining as much variation presented in the data set as possible. The use of a PC-ANN technique provided a good prediction of therapeutic indication of these herbal formulae as well as distinguishing these formulae from the one without therapeutic indication.

Key Words: Artificial neural network; Principal component analysis; Thai herbal formula

Introduction

Natural resources are plentiful in Thailand, and the country has its own traditional herbal medicine. Historically, the use of traditional herbal medicine has been a part of life since Sukhothai Period (1238 - 1377) (Subcharoen and Chuthaputti, 2006). With the entrance of western medicine, the role of traditional herbal medicine was declined due to lacking of scientific basis. However, high cost of new drugs, increased side effects and microbial resistance are some reasons for renewed interest in traditional herbal medicine. The development of traditional and herbal remedies supported by Thai government policy was launched through the implementation of the Fourth National Economic and Social Development Plan (1977-1981). Some items of herbal preparations, which have been used traditionally and widely by the people from time immemorial, were placed on the National List of Essential Medicines (Herbal Medicine List) and the National Traditional Household Remedies as part of an effort to promote the use of herbal preparations and provide diversity of alternatives for health care since 1999 (Chokevivat et al., 2005). Hence, the public interest in therapy based on traditional herbal medicine has been growing and the increasing effort has been directed towards scientific proof, clinical evaluation, and recipe analysis. Nowadays,

artificial intelligence (AI) has become an important field of study with a wide spread of applications in several fields. It also has been successfully applied to traditional medicine research area (Chen et al., 2006; Ung et al., 2007; Cao et al., 2009). In AI field, one of the most used branches is artificial neural networks (ANN). ANN is a computational model or mathematical model inspired in the natural biological neural networks (Zupan and Gasteiger, 1999). ANN method can simulate learning and generalize behavior of the human brain through data modeling and pattern recognition. The difference between an ANN model and a statistical model is that the ANN allows one to estimate relationships between one or several input variables called independent variables and one or several output variables called dependent variables without a specific mathematical function. Hence, an ANN works well for solving complicated non-linear problems of multivariate systems.

In this work, we used ANN to determine the appropriate model for the prediction of therapeutic indication of herbal formulae. Since the ratio between samples and variables in the ANN should be kept as high as possible, the principal component analysis (PCA) was used to trim down the number of variables in a sample data matrix. The combined use of PCA and ANN usually also improves the training speed, enhances the robustness of the model and reduces model errors.

Materials and Methods

Apparatus and Software

All input datasets were formatted as CSV files and stored for analysis by Microsoft Excel 2007. These data were processed by Intel® Core™ i3 computer having 2GB for RAM (Windows 7 operating system). The principal component analysis was performed by Scilab version 5.3.1 and artificial neural networks were implemented in WEKA version 3.7.5 using multilayer perceptron algorithm. Both programs are open source software.

Data Sources and Datasets

A set of 71 (47+24) herbal formulae official in the National List of Essential Medicine (NLEM) 2011 (the current version) and the National Traditional Household Remedies (NTHR) and 19 formulae without therapeutic indication (non-therapeutic formulae, NF) was used as a data set. Formulae without therapeutic indication were constructed by random combination of herbal items and by modification of existing herbal formulae. The 71 herbal formulae were classified into 7 categories based on therapeutic indication, e.g. cardiovascular system (CS, 10 recipes), gastrointestinal tract (GI, 23 recipes), obstetrics and gynecology (OG, 8 recipes), antipyretic (AP, 11 recipes), respiratory system (RS, 9 recipes), bone and muscle (BM, 8 recipes), and tonic (TN, 2 recipes). The list of 71 herbal formulae in NLEM and NTHR is shown in Table 1. The total number of Thai herbal items is 199. The data set of 90 herbal formulae was randomly divided into a training set (57 recipes), a monitoring set (10 recipes) and a validation set (23 recipes). A monitoring set was used to optimize the model parameters and a validation set was employed for testing the accuracy and precision of the models.

To make a standard for the amounts of herbs in the preparations, the amount of each herb in all preparations was normalized as the percentage of each herb to all herbs in the same preparations.

Results and Discussion

Muti-herb Formulae

The average number of herbs in formula for CS, GI, OG, AP, RS, BM and TN groups were 38.5, 16.5, 11.8, 12.8, 8.6, 10.1 and 6.0, respectively. The distribution of constituent herbs, in 71 herbal formulae, used in data set is illustrated in Figure 1. Most of these herbal formulae contain four or twelve herbal items.

Table 1 List of 71 herbal formulae selected from NLEM and NTHR

Thai herbal formulae	Therapeutic indication categories	Data sources	Thai herbal formulae	Therapeutic indication categories	Data sources
Ya-hom Tip-osod 1	CS	NLEM	Ya-Prasaplai 1	OG	NLEM
Ya-hom Tip-osod 2	CS	NTHR	Ya-Prasaplai 2	OG	NTHR
Ya-hom Tapepajit 1	CS	NLEM	Ya-Faipralaigun 1	OG	NLEM
Ya-hom Tapepajit 2	CS	NTHR	Ya-Faipralaigun 2	OG	NTHR
Ya-hom Navagote 1	CS	NLEM	Ya-Faihaagong 1	OG	NLEM
Ya-hom Navagote 2	CS	NTHR	Ya-Faihaagong 2	OG	NTHR
Ya-hom Intachak 1	CS	NLEM	Ya-Luadngam	OG	NLEM
Ya-hom Intachak 2	CS	NTHR	Ya-Satreelungklod	OG	NLEM
Ya-hom Kaelomwingwian	CS	NLEM	Ya-Kiewhom 1	AP	NLEM
Ya-Bumrung-lohit	CS	NTHR	Ya-Kiewhom 2	AP	NTHR
Ya-thad Bunjob 1	GI	NLEM	Ya-Chantaleela 1	AP	NLEM
Ya-thad Bunjob 2	GI	NTHR	Ya-Chantaleela 2	AP	NTHR
Ya-thad Obchoey	GI	NLEM	Ya-Prasachandaeng 1	AP	NLEM
Ya-Prasakaprao 1	GI	NLEM	Ya-Prasachandaeng 2	AP	NTHR
Ya-Prasakaprao 2	GI	NTHR	Ya-Prasaproayai 1	AP	NLEM
Ya-Prasakanplu 1	GI	NLEM	Ya-Prasaproayai 2	AP	NTHR
Ya-Prasakanplu 2	GI	NTHR	Ya-Mahanintangtong 1	AP	NLEM
Ya-Prasajetpangkee 1	GI	NLEM	Ya-Mahanintangtong 2	AP	NTHR
Ya-Prasajetpangkee 2	GI	NTHR	Ya-Haaraag	AP	NLEM
Ya-Muntathad 1	GI	NLEM	Ya-Ammareukwatee 1	RS	NLEM
Ya-Muntathad 2	GI	NTHR	Ya-Ammareukwatee 2	RS	NTHR
Ya-Wisaampayayai 1	GI	NLEM	Ya-Prasamawaeng 1	RS	NLEM
Ya-Wisaampayayai 2	GI	NTHR	Ya-Prasamawaeng 2	RS	NTHR
Ya-Mahachakyai 1	GI	NLEM	Ya-kae-ai-Pasomkanplu	RS	NLEM
Ya-Mahachakyai 2	GI	NTHR	Ya-kae-ai-Pasommanowdong	RS	NLEM
Ya-Luangpidsamut 1	GI	NLEM	Ya-kae-ai-Peunbaan-E-san	RS	NLEM
Ya-Luangpidsamut 2	GI	NTHR	Ya-Treepala	RS	NLEM
Ya-Apaisalee	GI	NLEM	Ya-Prabchompootaweep	RS	NLEM
Yatye-deekluafarang	GI	NLEM	Ya-Kasaisaen	BM	NLEM
Ya-Pasompetsangkart 1	GI	NLEM	Ya-kaelom-Ammappruek	BM	NLEM
Ya-Pasompetsangkart 2	GI	NLEM	Ya-Pasomkoklan 1	BM	NLEM
Ya-Ridseeduangmahakan	GI	NLEM	Ya-Pasomkoklan 2	BM	NLEM
Ya-tye	GI	NTHR	Ya-Pasomkoklan 3	BM	NLEM
Ya-Treegaysornmas	TN	NLEM	Ya-pasom-Taowanpriang 1	BM	NLEM
Ya-Treepigut	TN	NLEM	Ya-pasom-Taowanpriang 2	BM	NLEM
			Ya-Sahadtara	BM	NLEM

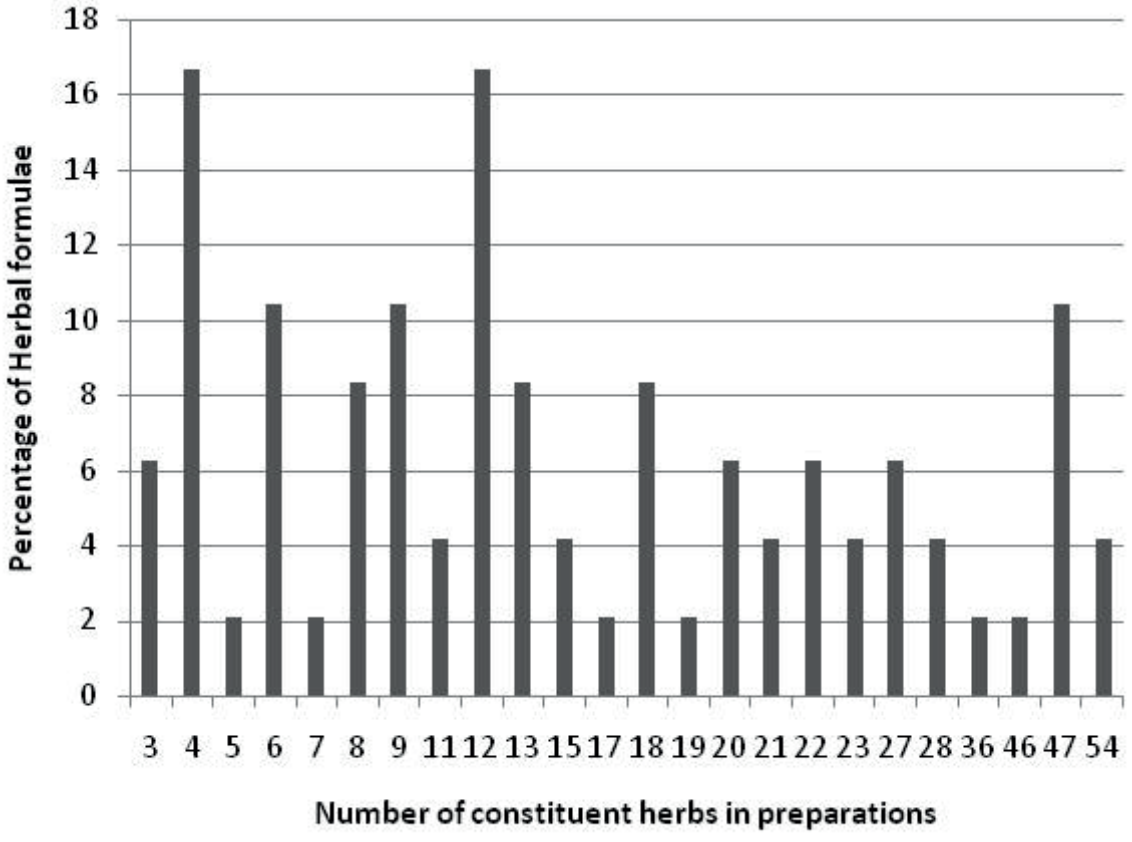

Figure 1 Distribution of herbal formulae with respect to the number of constituent herbs

Dimension Reduction with PCA

The performance of an ANN usually depends on data representation and one important characteristic of data representation is not interrelated. This is because correlated data reduce the distinctiveness of data representation and introduce confusion to the ANN model during the learning process (Mohamad-Saleh and Hoyle, 2008). The eliminating correlation in the sample data before they are submitted into an ANN is necessary. In this paper, an input data set consists of 90x199 (formulae x herb items). PCA was done onto this input data set prior to the ANN training process.

PCA technique is one of the multivariate data analysis approach based on the latent variable decomposition and is widely used for dimension reduction. PCA was first introduced by Pearson (1901), and independently developed by Hotelling (1933) (Jolliffe, 2002). This technique can compress all of the variables in which they are correlated, into a few principal components (PC), which are ordered by decreasing variability. The last of these components can be removed with minimum loss of real data. Thus, dimension of a sample data set can be reduced. The first PC defines the combination of variables that explains the greatest amount of variation and the second PC (independent to the first PC) indicates the next largest amount of variation and so on. The new variables, which are uncorrelated and called the principal components, are formed by taking linear combinations of the original variables. The principal component can be written as

$$z_{ij} = a_{i1}x_{1j} + a_{i2}x_{2j} + \ldots + a_{im}x_{mj} \qquad (1)$$

where z is the component score, a is the component loading, x is the measured value of variable, i is the principal component number, j is the sample number and m is the total number of variables (Cornish, 2007).

In applying PCA, the input data matrix was reduced as loading and score matrices were gathered. The first 49 components were observed to represent 98.23% of total variance. Once explored the data by PCA, a classification model was performed by the ANN into the next step.

ANN Modeling

In this application, the set of 57 herbal formulae (with 49 components) was employed as a training set and the ANN model was established by WEKA program with "multilayerPerceptron" algorithm in "classify" tab. This ANN model consisted of three layers of neurons, which were the basic computing units: the input layer with a number of active neurons corresponding to PC, one hidden layer with a number of active neurons, and the output layer with eight active neurons corresponding to the categories of therapeutic indication. The neurons were fully connected in a hierarchical manner. i.e. the outputs of one layer of nodes were used as inputs for the next layer and so on. The nodes in the input layer transfer the input data to all nodes in hidden layer. These nodes calculate a weighed sum of the inputs that is subsequently subjected to a non-linear transformation:

$$o_j = f[\sum_{i=1}^{I}(s_i w_{ij})] \quad (2)$$

where s_i is the input to node i in the input layer, I is the number of nodes in the input layer, w_{ij} (weights) are the connections between each node i in the input layer and each node j in hidden layer, and o_j is the output of node j in hidden layer, and f is a non-linear function called log-sigmoid function (Eq. (3)).

$$f(x) = \frac{1}{1+e^{-x}} \quad (3)$$

The log-sigmoid hidden layer is critical as it allows the network to learn non-linear relationships between inputs and outputs. The learning process was carried out through the back-propagation algorithm. The back-propagation network learns by calculating an error between desired and actual output and propagating this error information back to each node in the network. This back-propagation error is used to drive the learning at each node. The process of changing the weight of the connections to achieve some desired result is called learning or adaptation (Daniel et al., 1997).

To optimize ANN parameters (the number of neurons in the hidden layer, momentum and learning rate), the 10 recipes were constructed and used as monitoring set and all parameters were optimized by the Change One Separate factor at a Time (COST) technique. At this point, the mean square error (MSE) was calculated. Each time a new node was added to the hidden layer at arbitrary learning rate, momentum and the number of iterations. The number of neurons at the hidden layer, which had the minimum MSE value, was selected as the optimum number. After this step, the learning rate was varied from 0.1 to 0.9 at the optimum number of neurons at the hidden layer, arbitrary momentum and the number of iterations. The learning rate, which had the minimum MSE value, was selected as the optimum number. In the same way, the optimum number of momentum was defined. In order to avoid the overtraining and choosing the suitable number of epochs, the network was terminated before it learned idiosyncrasies present in the training data by searching the minimum MSE for the monitoring set. Finally, the number of the neurons at the hidden layer with the use of optimized momentum and learning rate was determined. Figure 2 illustrates the MSE value of the network with different learning rate and momentum and Figure 3 shows the variation

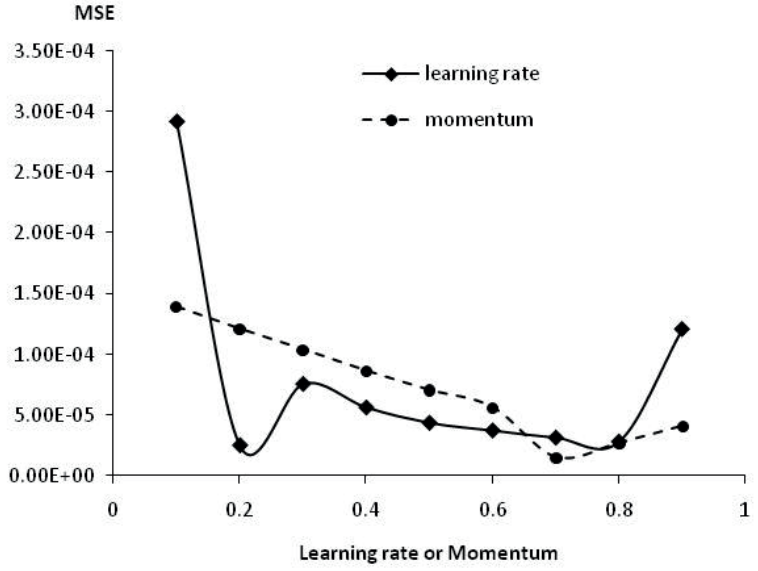

Figure 2 The relationship between learning rate (—) and momentum (.....) versus MSE

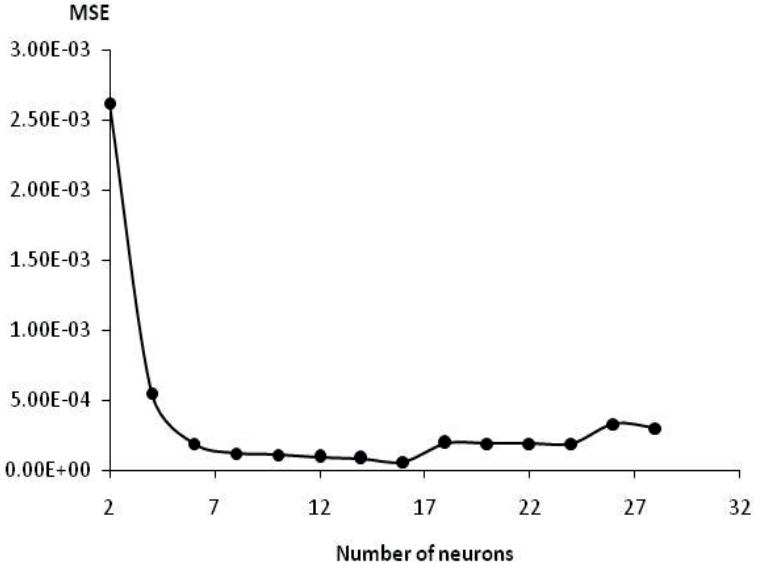

Figure 3 The relationship between number of neurons in hidden layer versus MSE

of MSE values of the network when the numbers of neurons in hidden layer was changed. The summary specifications for the network created for ANN models were listed in Table 2.

To study the ability of established ANN in the prediction of therapeutic indication for herbal formulae, 23 test herbal formulae were analyzed using the proposed method. The predicted class in each preparation was then compared with the known class in the respective preparation. The predicted results of ANN can be visualized by a confusion matrix (Table 3) of the desired and predicted

network output and by the performance indexes of the network.

A confusion matrix is a table that displays information about actual classification (target output) on the rows and predicted classification (network output) on the columns. The ideal classification result is to have large numbers down the main diagonal and small, ideally zero, off-diagonal elements of the matrix (Witten et al., 2011). The confusion matrix in Table 3 shows perfect classification for ANN model which the percentage of "correctly classified" of models is 100.00%. Furthermore, the performance indexes of such model was evaluated using data presented in the matrix and defined in terms of True Positive (TP) rate, False Positive (FP) rate and the percentage of accuracy. TP rate is the proportion of positive cases that are correctly classified, FP rate is the proportion of negative cases that are incorrectly classified, and the percentage of accuracy is the percentage of total number of predictions that are correct. The performance is confirmed by TP of 1, FP of 0 and the percentage of accuracy of 100.00%. This study indicates that combination of herbs and herbal proportions in Thai herbal formulae were rationally formulated and structurally defined. However, this ANN model has some limitation in that it is used to predict class of therapeutic indications of herbal formulae with herbal constituents modified around those from 71 formulae in NLEM or NTHR.

Table 2 Artificial neural network specifications and parameters

Parameter	ANN
Input nodes	49
Hidden nodes	16
Output nodes	8
Learning rate	0.2
Momentum	0.7
Hidden layer transfer function	log-sigmoid
Optimum number of iterations	4000

Conclusion

The capability of a hybrid methodology that merges PCA techniques and an ANN modeling technique is presented and evaluated in this paper. In accordance with the achieved results, the suggested model that used PC as input variables could predict the therapeutic indication group. The performance

Table 3 A confusion matrix for 23 herbal formulae when applying ANN model to the validation set

		Network Output							
		CS	GI	OG	AP	RS	BM	TN	NF
Target Output	CS	4	0	0	0	0	0	0	0
	GI	0	5	0	0	0	0	0	0
	OG	0	0	2	0	0	0	0	0
	AP	0	0	0	4	0	0	0	0
	RS	0	0	0	0	1	0	0	0
	BM	0	0	0	0	0	1	0	0
	TN	0	0	0	0	0	0	1	0
	NF	0	0	0	0	0	0	0	5

of this model was improved with extraction of the factors that poorly contributed in the therapeutic indications and simplified the training procedure of the ANN by the inclusion of only the significant PC in the model. Furthermore, the PC-ANN model is useful for not only predicting class of therapeutic indications but also validating herbal formulae, distinguishing the formulae from the formulae without therapeutic indication.

References

Cao, T., Kamei, K., and Dang T. L. (2009) Visualization System of Herbal Prescription Effects in Oriental Medicine by Self-Organizing Map. *Biomedical Soft Computing and Human Sciences* 14(1): 101-108.

Chen, X., et al. (2006) Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation. *British Journal of Pharmacology* 149: 1092-1103.

Chokevivat, V., Chuthaputti, A. and Khumtrakul, P. (2005) The Use of Traditional Medicine in the Thai Health Care System. In *Regional Consultation on Development of Traditional Medicine in the South East Asia Region* (World Health Organization Regional Office for South East Asia), Document no.9. Pyongyang, DPR Korea.

Cornish, R. (2007) Principal Component Analysis. In *Statistics*, [Online URL: www.mlsc.lboro.ac.uk/resources/statistics/3.2PrincipleCom.pdf] accessed on September 24, 2012.

Daniel, S., Kvasnicka, V., and Pospichal, J. (1997) Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems* 30: 43-62.

Jolliffe, I. T. (2002) *Pricipal Component Analysis* 2nd ed., Springer-Verlag, USA.

Mohamad-Saleh, J. and Hoyle, B. S. (2008) Improved Neural Network Performance Using Principal Component Analysis on Matlab. *International Journal of The Computer, the Internet and Management*, 16(2): 1-8.

Subcharoen, P. and Chuthaputti, A. (2006) Thai Traditional Medicine Kingdom of Thailand. In *WHO Global atlas of traditional, complementary and alternative medicine*, pp.103-106. The WHO Centre for Health Development, Kobe.

Ung, C. Y., et al. (2007) Are herb-pairs of traditional Chinese medicine distinguishable from others? Pattern analysis and artificial intelligence classification study of traditionally defined herbal properties. *Journal of Ethopharmacology* 111: 371-377.

Witten, I. H., Frank, E., and Hall, M. A. (2011) *Data Mining Practical Machine Learning Tools and Techniques* 3rd ed., pp.164. Morgan Kaufmann Publishers, Elsevier, USA.

Zupan, J. and Gasteiger, J. (1999) *Neural Networks in Chemistry and Drug Design*, 2nd ed., Wiley-VCH Publishing Company, Weinheim.