DOI: 10.1002/cmdc.200600166

Virtual Screening for PPAR Modulators Using a Probabilistic Neural Network

Swetlana Derksen,[a] Oliver Rau,[b] Petra Schneider,[c] Manfred Schubert-Zsilavecz,[b] and Gisbert Schneider*[a]

Probabilistic neural networks (PNNs) were originally conceived by Specht in 1988 as a method for estimating probability density functions.[1–3] Although PNNs have existed for almost two decades, they have found only sporadic applications in drug discovery as classifiers for the prediction of pharmacological molecular properties.[4–6] Herein we apply PNNs for the first time in a prospective virtual screening study.

The basic idea of PNNs is to use classes of training data (such as compounds with known pharmacological activity) as reference points and to determine the probability of each new data point (test compound) belonging to one of the training data categories. These categories could be expressed as "active" and "inactive", or represent substances that bind to individual receptor subtypes. As a result of PNN calculation, a probability is assigned to each test compound with which it falls into one of the data categories. In our study, PNNs were used to predict peroxisome proliferator-activated receptor (PPAR) modulators, in particular PPARα and PPARγ agonists. PPARs are nuclear receptors comprising three subtypes: PPARα (NR1C1), PPARδ (NR1C2), and PPARγ (NR1C3). After activation by an agonist, the PPARs function as transcription factors for enzymes and regulatory proteins involved in the primary metabolism of glucose and fatty acids.[7] Synthetic agonists for PPARα and PPARγ, represented by fibrate and glitazone-type drugs, are used clinically in the treatment of dyslipidemias and type 2 diabetes.[8–12] Current research focuses on dual PPARα/γ agonists, as well as the potential of PPAR agonists in general, for the treatment of diseases involving inflammatory processes.[13,14]

[a] S. Derksen, Prof. Dr. G. Schneider
Johann Wolfgang Goethe University
Institute of Organic Chemistry and Chemical Biology/ZAFES
Siesmayerstrasse 70, 60323 Frankfurt/Main (Germany)
Fax: (+49) 69-798-24880
E-mail: g.schneider@chemie.uni-frankfurt.de

[b] O. Rau, Prof. Dr. M. Schubert-Zsilavecz
Johann Wolfgang Goethe University
Institute of Pharmaceutical Chemistry/ZAFES
Max-von-Laue Strasse 9, 60438 Frankfurt/Main (Germany)

[c] Dr. P. Schneider
Schneider Consulting GbR
George-C.-Marshall Ring 33, 61440 Oberursel (Germany)

Supporting information for this article is available on the WWW under http://www.chemmedchem.org or from the author.

The Specs catalogue[1] provided the compound pool for virtual screening using the PNN technique. It contained 229658 small organic molecules that can be purchased to build up screening libraries. From this pool, the PNN predicted potential PPAR ligands, and nine top-scoring substances were tested for PPAR agonistic modulation in a cellular reporter-gene assay.

Data compilation: All molecules for PNN development were extracted from the COBRA collection.[15] This database is a compilation of leads, lead candidates, and drugs covering all major receptor classes with a focus on recently developed agents. COBRA v.3.5 contained 5693 molecules, including 69 ligands that bind to PPAR nuclear hormone receptors (see Supporting Information); among these, 51 are selective PPAR modulators (12 for PPARα, 34 for PPARγ, 5 for PPARδ) and 18 represent multiple modulators (15 dual modulators of PPARα and PPARγ, one dual modulator of PPARα and PPARδ, and two triple modulators). Prior to virtual screening with PNNs, we desalted and neutralized all compounds using our own SVL script within the modeling suite MOE[2].

Molecular descriptor: Each molecule was encoded by the standard 150-dimensional CATS topological pharmacophore descriptor.[16,17] CATS descriptors consider all possible pairs of five generic atom types (lipophilic, hydrogen-bond donor, hydrogen-bond acceptor, positively charged or polarizable, negatively charged or polarizable) spaced at intervals of zero to nine bonds and their scaled frequency of occurrence. All methods were implemented in the programming language Java,[3] involving the CDK-Package[4] and JAMA[5].

Network training: PNNs can be considered as a method for similarity assessment. Test data (screening compounds) are compared with training data (reference compounds), and the test data points are assigned the category of the training data points that are closest to the test data. The neighborhood definition is managed by applying a Gaussian function based on the idea of a "Parzen window":[18] the principle of PNN training is to place a Gaussian function at the location of each training data point, and to form the probability density function for each data category by summing over those Gaussians that belong to the data points of each category (Figure 1). The standard deviation σ specifies the "width" of the Gaussian window and corresponds to the surrounding of a pattern. It has been stated that PNNs do not require time-consuming training because Gaussians are simply placed on the training data points.[5] This is only true if the standard deviations are not subjected to an optimization procedure. In our studies with PNNs, we generally found that optimization of the widths of the Gaussian functions for each data category was beneficial for PNN performance.

Formally, the *a posteriori* probability of a molecule x belonging to category k is computed by a PNN according to Equation (1):

[1] Version: June 2003; Specs, HT Delft, The Netherlands; http://www.specs.net
[2] Chemical Computing Group, Montreal, Canada; http://www.chemcomp.com
[3] http://www.java.sun.com
[4] http://cdk.sf.net
[5] http://math.nist.gov/javanumerics/jama

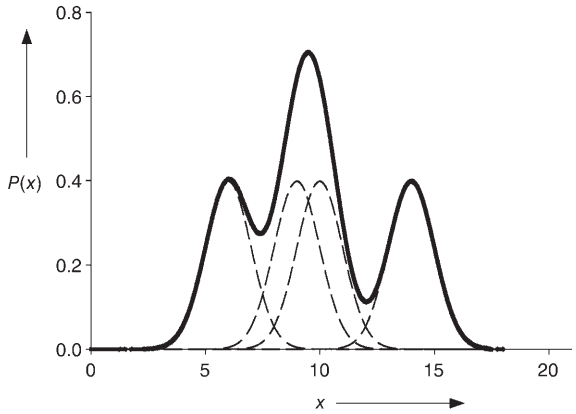

Figure 1. Principle of PNN training and prediction. The neighborhood of each training data point is captured by a Gaussian centered at the respective descriptor values *x*. In this example, there are four data points captured by univariate Gaussians (dotted lines). The bold line indicates the resulting Gaussian value for this data class.

$$g_k(x) = \frac{1}{2\pi^{d/2}\sigma_k^d M_k} \sum_{j=1}^{M_k} \exp^{K(x,y_j)} \quad (1)$$

in which *y* are the training data, *d* the number of molecular descriptors, and *M* the number of data points representing the data category ("active" or "inactive").

Each training data point contributes a probability that the test data point was generated by a Gaussian centered on the associated training point. The contribution is computed by the Kernel function *K*, which represents the distance between the training data point *y* and the test data point *x* [Eq. (2)]:

$$K(x,y) = \frac{\sum_i x_i y_i - 1}{\sigma_k^2} \quad (2)$$

for which *i* is the i^{th} descriptor value.

Prior to PNN computing, the descriptors of a data point were scaled so that $\sum_i x_i^2 = 1$. Standard deviations were optimized for each data category by using a (μ,λ) evolution strategy,[19,20] with $\mu = 1$ parent, and $\lambda = 50$ offspring per generation. The evolution strategy is an adaptive stochastic search method following iterative evaluation of populations of possible solutions. Each parameter variation corresponds to an individual. Fitness was determined by evaluating the PNN prediction accuracy for each individual [Eq. (3)]. Optimization was performed with a fixed number of 100 generations.

For an estimate of prediction accuracy, multiple cross-validation of the PNN model was performed. First, all data were divided into three equally sized sets. Then the following procedure was repeated three times, each time leaving out one of the three initial data sets as validation data: the remaining two sets were pooled again and provided the basis for network training and testing (80 + 20 split into training and test data). Training data were used to build up the PNN, and test data were used to evaluate the prediction accuracy during parameter optimization with the evolution strategy. The split into training and test sets and subsequent optimization were repeated 10 times. After network optimization, the prediction accuracy was estimated by using the validation data. Results from repeated optimization runs were averaged to obtain the mean error. The final PNN for compound selection was trained with all available data.

PNN accuracy was quantified by Matthews' correlation coefficient (*MCC*) for a two-class prediction [Eq. (3)].

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TN+FN)(TN+FP)(TP+FN)(TP+FP)}} \quad (3)$$

for which *TP*, *TN*, *FP*, and *FN* are the numbers of true positive, true negative, false positive, and false negative predictions, respectively. *MCC* values can range from −1 to 1, where 1 indicates perfect classification.[21]

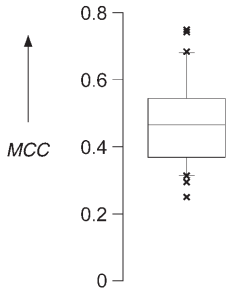

Figure 2. Cross-validated prediction accuracy of a PNN trained on 69 PPAR and 5693 COBRA compounds using CATS2D descriptors. The box shows the median with 25% and 75% percentiles; whiskers give the 5% and 95% percentiles; outliers are indicated by asterisks. *MCC*: Matthews' correlation coefficient [Eq. 3].

Results and discussion: A PNN was trained separating the two categories "PPAR modulators" and "others" using 69 PPAR reference ligands and 5693 "other" compounds. An average accuracy of prediction of *MCC* = 0.47 was obtained (Figure 2). The σ values of the Gaussians were 0.11 for "PPAR modulators" and 0.33 for "others". This indicates that PPAR modulators form isolated chemotype "islands" in chemical space. We also observed that prediction accuracy differed for the individual cross-validation sets. This was probably caused by the small number and heterogeneity of the PPAR ligand structures. To analyze this in more detail, we constructed a self-organizing map (SOM) projection of the data.[22,23] The projection supports the idea of having separate clusters of PPAR modulators (not shown). This means that the PNN had to capture potentially different structure–activity relationships, one for each ligand class. A larger set of diverse training compounds would help improve the generalization performance of the PNNs, because each cluster would be better represented in the PNN and cover a part of the PPAR pharmacophore space. Training results also indicate that several of the "other" compounds are PPAR ligands, meaning that the training task was partially ill-posed. This interpretation is supported by the notion that PPAR can accommodate various types of ligands.[7]

Subsequent virtual screening of the Specs catalogue yielded nine molecules **1**–**9** (Figure 3) that were selected for testing of their ability to activate PPARα and PPARγ (Table 1). This final selection was done by visual inspection of the top-ranking 20 compounds. We picked different chemotypes with low calculated lipophilicity (Slog *P*[24] as implemented in MOE).

Activity was recorded as a relative activity against that of the reference compounds WY14,643 (PPARα) and pioglitazone

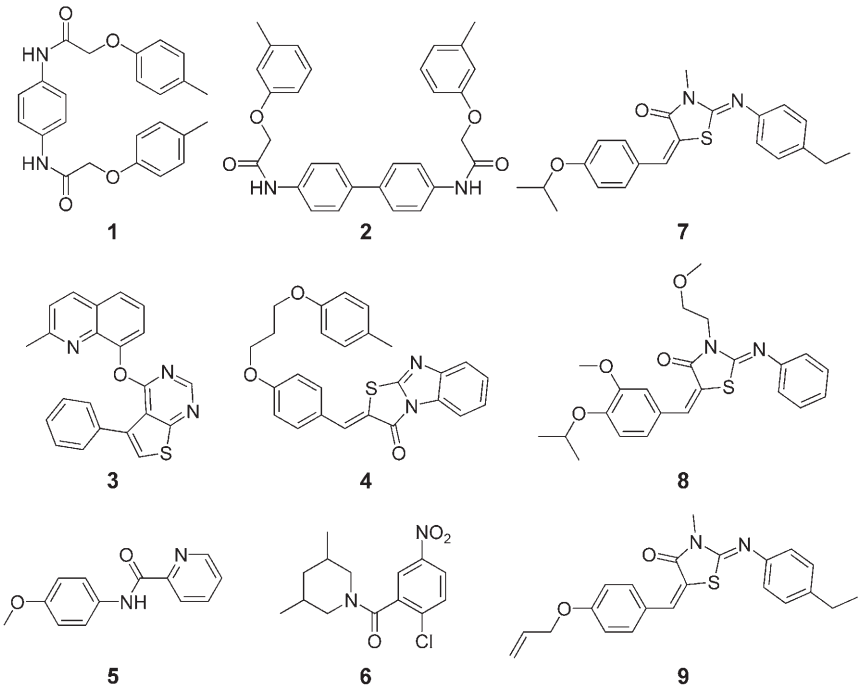

Figure 3. Molecules that were predicted to be PPAR modulators by the PNN and tested in a cell-based reporter gene assay.

Table 1. Results of in vitro pharmacological assays.[a]

Compound	PPARα	PPARγ
1	i.a.	i.a.
2	i.a.	i.a.
3	i.a.	i.a.
4	i.a.	i.a.
5	i.a.	i.a.
6	i.a.	0.52 ± 0.21
7	i.a.	0.54 ± 0.08
8	0.26 ± 0.05	0.17 ± 0.05
9	i.a.	0.14 ± 0.05

[a] Expressed as the mean relative activation (± SD, $n = 3$) of the targets at a compound concentration of 100 μM; i.a. = inactive.

(PPARγ). Compounds **6**–**9** exhibited weak agonist activity toward PPARγ; only compound **8** showed some additional activity for PPARα (Figure 4). The chloronitrobenzoic acid motif of compound **6** is present in the known PPARγ antagonists GW9662 (**10**)[25] and T0070907 (**11**)[26] (Figure 5). However, the amidic nitrogen atom in compounds **10** and **11** originates from an aniline and an aminopyridine, respectively, whereas in compound **6** the amidic nitrogen atom is part of a piperidine derivative. For T0070907, a covalent interaction with Cys 285 in helix 3 of PPARγ has been shown.[26]

Compounds **7**, **8**, and **9** share a motif that varies at only three positions: 1) N substitution at the thiazolidine group, 2) aliphatic derivatization of the aniline moiety, and 3) substitution pattern of the benzylidene group. These seemingly minor changes led to both a loss of PPARγ activation for compounds **8** and **9** relative to **7** and a gain of PPARα activity for compound **8**. The similarity of compounds **7**–**9** to the glitazone drugs pioglitazone (**12**, EC_{50} = 550 nM), and rosiglitazone (**13**, EC_{50} = 80 nM) is striking,[27] with the thiazolidinedione (TZD) building block containing a masked carbonyl functionality. In contrast to the glitazones, compounds **7**–**9** do not contain an acidic group, which is present in many PPARα and PPARγ agonistic drugs. Notably, the free TZD head group alone of **12** and **13** does not explain the strong receptor activation, as netoglitazone has a poorer in vitro potency toward PPARγ (EC_{50} = 8 μM),[28] but still contains the unsubstituted TZD group.

To get an idea of the potential binding mode of our most potent agonist **7**, we performed automated molecular docking

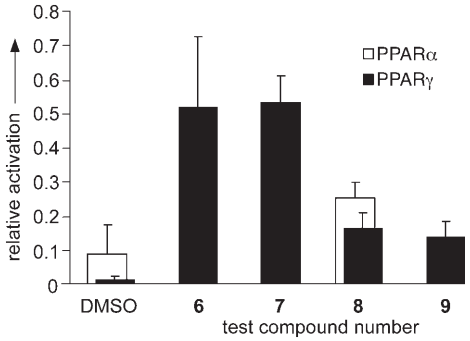

Figure 4. Transactivation of a luciferase reporter gene in Cos7 cells by the Gal4–PPAR chimeric receptor for the PPAR subtypes PPARα and PPARγ. The figure shows relative activation in comparison with the effect of the reference compound WY14,643 (100 μM) for PPARα and pioglitazone (**12**, 1 μM) for PPARγ. Bars represent the mean ($n = 3$) ± SD; all values are significantly different from DMSO negative control, with $P > 95\%$.

into a PPARγ complex structure (PDB code 2F4B) with the software GOLD.[29,30] We chose this receptor structure because it contains a large indole-based ligand **14** (PPARγ: EC_{50} = 70 nM; PPARα: EC_{50} = 8 nM) that lacks the TZD group[31] and thus provides a reference structure for the comparably large binding pocket of PPARγ (≈1400 Å^3 in apo-PPARγ)[32] without providing a preformed TZD binding moiety. For the docking experiment, the binding pocket was defined by a radius of 5 Å around ligand **14**. Prior to docking of compound **7**, ligand **14** was removed from the receptor pocket. The average fitness (GoldScore function) of the 20 top-scoring docking solutions yielded a value of 51, indicating favorable complex formation. A preferred potential binding mode was detected which fills only part

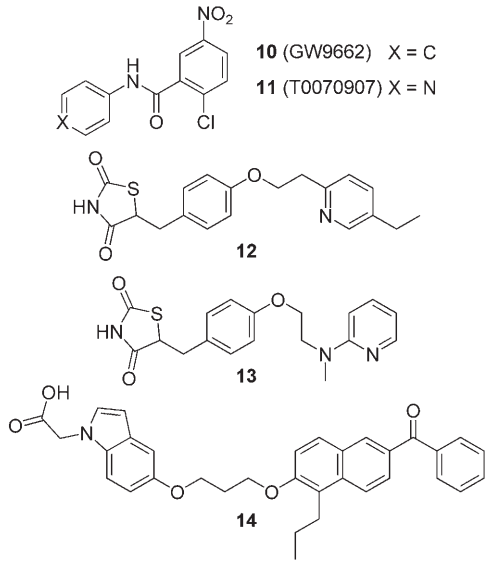

Figure 5. Known PPAR modulators.

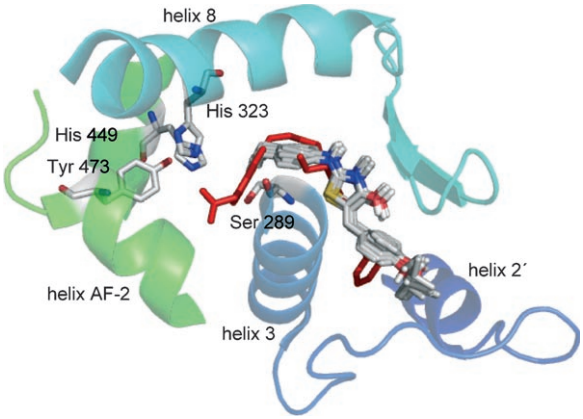

Figure 6. Experimental binding mode of the unselective but potent ligand **14** (red) to the ligand binding domain of PPARγ (PDB code 2F4B, resolution 2.3 Å). Crucial interaction points for receptor activation are highlighted. The predicted preferred orientation of ligand **7** is shown by the superposition of several high-ranking docking solutions.

of the receptor pocket (Figure 6). Studies with ligand **14** revealed that the distance between the acid group and the linker were important for receptor activation,[31] which is also supported by the experimentally determined binding mode of rosiglitazone **13** (PDB code 2PRG, not shown).[33] The docking solutions suggest, however, that ligand **7** does not form a U-shape around helix 3 and extend to the activation function helix AF-2 (Tyr 473), which could explain its comparably low potency. According to our simulation, direct interaction with the helix AF-2 is not strictly required for partial PPARγ activation. This interpretation is corroborated by a novel class of selective PPARγ partial agonists recently discovered by Lu and co-workers using shape-based similarity searching and subsequent molecular docking.[34] These pyrazol-5-ylbenzenesulfonamide-based compounds do not form hydrogen bonds to helix AF-2 but interact primarily with helix 3.

Whereas other nuclear receptor ligand binding cavities are nearly completely filled with their agonist, that of PPAR provides extra space, which can be occupied by different kinds of agonists. Unsaturated fatty acids, endogenous ligands of PPAR, predominantly bind to PPAR through lipophilic interactions, leading to the assumption that the PPAR ligand binding cavity can accommodate a variety of lipophilic compounds.[7] The list of candidate compounds produced by our PNN virtual screening mirrors this ligand diversity. Notably, many of these compounds are able to form a U-shape—a structural feature of many PPAR agonists that has been implicitly captured by the PNN model.

Conclusions: PNNs were successful in a pilot virtual screening study. Their prediction accuracy was not perfect, but provided a meaningful separation of PPAR modulators and other drugs. For widely spread data that forms many small clusters, it was difficult to establish a predictive PNN model by optimizing one standard deviation value per category. The PPAR-versus-drug discriminative model was effectively trained based on topological pharmacophore (CATS2D) descriptors. The molecules that were retrieved from the Specs compound library by PNN prediction possess features that are present in known PPAR ligands. The PNN system was able to retrieve PPARγ agonists from a vendor compound collection that likely do not show the glitazone-type of receptor–ligand interaction through helix AF-2 of PPAR. Our study also suggests that crucial features for receptor activation are still missing in these molecules. The ligand docking experiment provides a theoretical basis for future molecular design, which could either aim at the interaction of variants of ligand **7** with helix AF-2 of PPARγ to obtain full agonists, or yield a different binding mode of partial agonists.[34] We conclude that the PNN approach is suited for parallel similarity searching with multiple reference ligands and the construction of focused compound libraries.

Experimental Section

Reporter gene assay: The assay was carried out as described previously.[35,36] Cos7 cells were seeded at 30 000 cells per well in a 96-well plate and transfected the following day with pFR-Luc (Stratagene) containing a luciferase gene downstream of a Gal4 response element, the expression plasmid pFA-CMV-PPAR for the respective PPAR subtype, and pRL-SV40 (Promega) for normalization, by using Lipofectamine2000 (Invitrogen). Four hours after transfection, cells were incubated in triplicate wells with test compounds (100 μM) overnight at a final DMSO concentration of 0.1% *v/v*. Determination of reporter gene activity and normalization for transfection efficacy and cell number was carried out with DualGlo (Promega) and a GENiosPro luminometer (Tecan). Each assay was repeated three times, and each value shown is significantly different from DMSO negative control, with $P > 95\%$ (one-sided *t* test).

Docking experiments: The docking software GOLD (version of 2005, URL: www.ccdc.cam.ac.uk/products/life_sciences/gold/)[29,30] was used with the GoldScore fitness function and the following settings. Population: maxops = 100 000, popsiz = 100, select_pressure = 1.1, n_islands = 5, niche_siz = 2; Genetic Algorithm: pt_

crosswt = 95, allele_mutatewt = 95, migratewt = 10; Search: start_vdw_linear_cutoff = 4.0, initial_virtual_pt_match_max = 2.5. Scaling of internal vdW energy: log scale from 0.2 to 1 over 75000 ops, sampling interval (operations) = 1000; scaling of virtual pt maximum length: linear scale from 2.5 to 2 over 75,000 ops; sampling interval (operations) = 3000; scaling of 4–8 potential cutoff log scale from 4 to 120 over 75,000 ops, sampling interval (operations) = 3000.

Acknowledgements

Norbert Dichter is warmly thanked for technical assistance. This research was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Frankfurt am Main.

Keywords: computer chemistry · drug design · machine learning · PPAR agonists · reporter gene assays

[1] D. F. Specht, *Proc. IEE International Conference on Neural Networks* **1988**, pp. 525–532.
[2] T. M. Cover, J. A. Thomas, *Elements of Information Theory*, Wiley, New York, **1991**.
[3] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, 2nd ed., Wiley, New York, **2001**.
[4] T. Niwa, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 113 – 119.
[5] T. Niwa, *J. Med. Chem.* **2004**, *47*, 2645 – 2650.
[6] S. Vilar, L. Santana, E. Uriarte, *J. Med. Chem.* **2006**, *49*, 1118 – 1124.
[7] B. Desvergne, W. Wahli, *Endocr. Rev.* **1999**, *20*, 649 – 688.
[8] H. Hauner, *Diabetes/Metab. Res. Rev.* **2002**, *18* Suppl. 2, S10 – S15.
[9] S. E. Inzucchi, *J. Am. Med. Assoc.* **2002**, *287*, 360 – 372.
[10] J. P. Berger, T. E. Akiyama, P. T. Meinke, *Trends Pharmacol. Sci.* **2005**, *26*, 244 – 251.
[11] B. Staels, W. Koenig, A. Habib, R. Merval, M. Lebret, I. P. Torra, P. Delerive, A. Fadel, G. Chinetti, J. C. Fruchart, J. Najib, J. Maclouf, A. Tedgui, *Nature* **1998**, *393*, 790 – 793.
[12] S. Kersten, B. Desvergne, W. Wahli, *Nature* **2000**, *405*, 421 – 424.
[13] A. Castrillo, P. Tontonoz, *Annu. Rev. Cell Dev. Biol.* **2004**, *20*, 455 – 480.
[14] R. Genolet, W. Wahli, L. Michalik, *Curr. Drug Targets: Inflammation Allergy* **2004**, *3*, 361 – 375.
[15] P. Schneider, G. Schneider, *QSAR Comb. Sci.* **2003**, *22*, 713 – 718.
[16] G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angew. Chem.* **1999**, *111*, 3068 – 3070; *Angew. Chem. Int. Ed.* **1999**, *38*, 2894 – 2896.
[17] U. Fechner, L. Franke, S. Renner, P. Schneider, G. Schneider, *J. Comput.-Aided Mol. Des.* **2003**, *17*, 687 – 698.
[18] E. Parzen, *Ann. Math. Stat.* **1962**, *33*, 1065 – 1076.
[19] I. Rechenberg, *Optimierung technischer System nach Prinzipien der biologischen Evolution*, Frommann-Holzboog, Stuttgart, **1973**.
[20] D. Whitley, *J. Inf. Softw. Tech.* **2001**, *43*, 817 – 831.
[21] B. W. Matthews, *Biochim. Biophys. Acta* **1975**, *405*, 442 – 451.
[22] T. Kohonen, *Self-organizing Maps*, Springer, Berlin, **2001**.
[23] T. Kohonen, *Biol. Cybern.* **1982**, *43*, 59 – 69.
[24] S. A. Wildman, G. M. Crippen, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868 – 873.
[25] J. T. Huang, J. S. Welch, M. Ricote, C. J. Binder, T. M. Willson, C. Kelly, J. L. Witztum, C. D. Funk, D. Conrad, C. K. Glass, *Nature* **1999**, *400*, 378 – 382.
[26] G. Lee, F. Elwood, J. McNally, J. Weiszmann, M. Lindstrom, K. Amaral, M. Nakamura, S. Miao, P. Cao, R. M. Learned, J. L. Chen, Y. Li, *J. Biol. Chem.* **2002**, *277*, 19649 – 19657.
[27] T. M. Willson, P. J. Brown, D. D. Sternbach, B. R. Henke, *J. Med. Chem.* **2000**, *43*, 527 – 550.
[28] B. R. Henke, S. G. Blanchard, M. F. Brackeen, K. K. Brown, J. E. Cobb, J. L. Collins, W. W. Harrington Jr., M. A. Hashim, E. A. Hull-Ryde, I. Kaldor, S. A. Kliewer, D. H. Lake, L. M. Leesnitzer, J. M. Lehmann, J. M. Lenhard, L. A. Orband-Miller, J. F. Miller, R. A. Mook Jr., S. A. Noble, W. Oliver Jr., D. J. Parks, K. D. Plunket, J. R. Szewczyk, T. M. Willson, *J. Med. Chem.* **1998**, *41*, 5020 – 5036.
[29] G. Jones, P. Willett, R. C. Glen, *J. Mol. Biol.* **1995**, *245*, 43 – 53.
[30] G. Jones, P. Willett, R. C. Glen, A. R. Leach, R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727 – 748.
[31] N. Mahindroo, C. C. Wang, C. C. Liao, C. F. Huang, L. L. Lu, T. W. Lien, Y. H. Peng, W. J. Huang, Y. T. Lin, M. C. Hsu, C. H. Lin, C. H. Tsai, J. T. Hsu, X. Chen, P. C. Lyu, Y. S. Chao, S. Y. Wu, H. P. Hsieh, *J. Med. Chem.* **2006**, *49*, 1212 – 1216.
[32] B. R. Henke, *Prog. Med. Chem.* **2004**, *42*, 1 – 53.
[33] R. T. Nolte, G. B. Wisely, S. Westin, J. E. Cobb, M. H. Lambert, R. Kurokawa, M. G. Rosenfeld, T. M. Willson, C. K. Glass, M. V. Milburn, *Nature* **1998**, *395*, 137 – 143.
[34] I. L. Lu, C. F. Huang, Y. H. Peng, Y. T. Lin, H. P. Hsieh, C. T. Chen, T. W. Lien, H. J. Lee, N. Mahindroo, E. Prakash, A. Yueh, H. Y. Chen, C. M. V. Goparaju, X. Chen, C. C. Liao, Y. S. Chao, J. T. A. Hsu, S. Y. Wu, *J. Med. Chem.* **2006**, *49*, 2703 – 2712.
[35] O. Rau, M. Wurglics, A. Paulke, J. Zitzkowski, N. Meindl, A. Bock, T. Dingermann, M. Abdel-Tawab, M. Schubert-Zsilavecz, M, *Planta Med.* **2006**, *72*, 881 – 887.
[36] S. Ulrich, M. Loitsch, O. Rau, M. Schubert-Zsilavecz, J. M. Stein, *Cancer Res.* **2006**, *66*, 7348 – 7354.

Received: July 9, 2006
Published online on October 26, 2006

