Applications of Self-Organizing Neural Networks in Virtual Screening and Diversity Selection[†]

Paul Selzer* and Peter Ertl

Novartis Institutes for BioMedical Research, Cheminformatics, CH-4002 Basel, Switzerland

Received February 28, 2006

Artificial neural networks provide a powerful technique for the analysis and modeling of nonlinear relationships between molecular structures and pharmacological activity. Many network types, including Kohonen and counterpropagation, also provide an intuitive method for the visual assessment of correspondence between the input and output data. This work shows how a combination of neural networks and radial distribution function molecular descriptors can be applied in various areas of industrial pharmaceutical research. These applications include the prediction of biological activity, the selection of screening candidates (cherry picking), and the extraction of representative subsets from large compound collections such as combinatorial libraries. The methods described have also been implemented as an easy-to-use Web tool, allowing chemists to perform interactive neural network experiments on the Novartis intranet.

1. INTRODUCTION

Modern pharmaceutical research would not be possible without the intensive application of cheminformatics.[1−3] The fact that consecutive phases of the drug discovery pipeline become successively more expensive means that it is highly desirable to exclude drug candidates with a low probability of success as early as possible from the development pipeline, while speeding up the development of more promising compounds. Cheminformatics-based methods can support this process by providing rational decision assistance, ranging from the effective organization of data in corporate data warehouses,[4] supporting fast structure searching, to the rationalization of relationships between molecular structure and biological activity and the prediction of molecular properties.

Pharmaceutical companies have the great advantage of having vast amounts of in-house chemical and related biological data from standardized assay protocols available for analysis and model building. Depending on the availability of data for a given project, cheminformatics can provide information in a synergistic manner that is complementary to other computational disciplines such as molecular modeling or bioinformatics.

There are many different statistical methods used in cheminformatics for the analysis of data. In this work, we want to focus on Kohonen and counterpropagation (CPG) neural networks and their contribution to the pharmaceutical research process. We will give examples of the most prominent and most important applications of these methods, including the clustering of chemical structures, prediction of molecular properties, rational selection of screening compounds, and the creation of representative subsets from large compound collections, for example, combinatorial libraries.

In section 3, we will describe how this method was implemented as a Web tool that can be used by scientists to run and analyze interactive neural network experiments on the Novartis intranet.

2. METHODOLOGY

Artificial neural networks (ANNs)[5−7] provide a powerful technique for modeling nonlinear relationships. Neural networks are, therefore, commonly applied in pharmaceutical research to analyze the complex relationships that exist between the structure of molecules and their physicochemical or biological properties, with the goal of identifying which structural features are of pharmacological importance.

For this purpose, we used molecular descriptors, calculated from intramolecular atomic distances in three-dimensional (3D) space, to describe the 3D shape of molecules. These descriptors allowed 3D structural information to be used as the input required for training the neural networks.

The neural network was used to cluster the compounds in a two-dimensional (2D) map on the basis of the similarity (or diversity) of their descriptors. Coloring each neuron according to the properties (e.g., physicochemical properties or biological activity) of its constituent compounds allowed the analysis of correspondence between the structural features and molecular properties of the compounds. This information can be used in the prediction of molecular properties, the selection of compounds for pharmacological screening, and the generation of representative subsets from large compound collections, as described in section 3.

2.1. Radial Distribution Function (RDF) as Structure Descriptor. Artificial neural networks require a fixed length vector representation of the input and the output training data. The 3D structure of a molecule, or more precisely, the spatial arrangement of its pharmacophore features, determines the pharmacological properties of the molecule. We therefore used RDF[8,9] molecular descriptors, which express pharmacophore features by coding the arrangement of atomic properties in 3D space as a vector of real numbers.

[†] Dedicated to Professor Johann Gasteiger.

* Corresponding author fax: +41616964069; e-mail: paul.selzer@novartis.com

10.1021/ci0600657 CCC: $33.50
Published on Web 07/21/2006

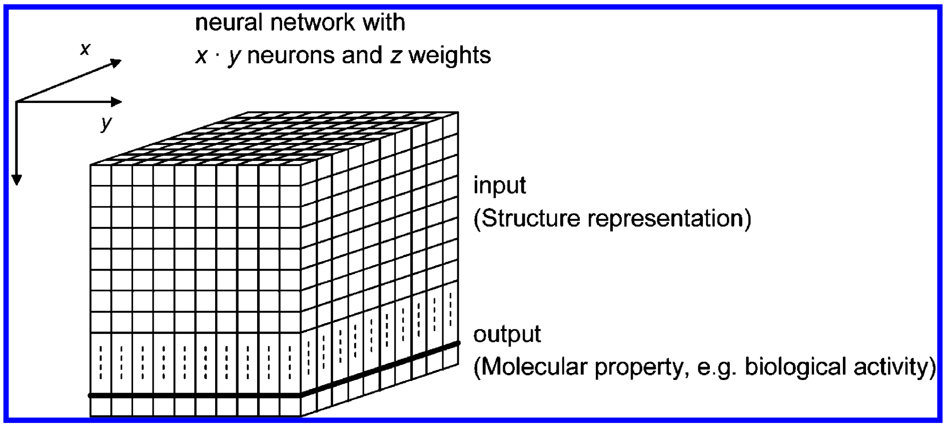

neural network with
x · y neurons and z weights
x
y
input
(Structure representation)
output
(Molecular property, e.g. biological activity)

Figure 1. Neural network of type counterpropagation (CPG), which consists of a two-dimensional arrangement of neurons. Each neuron has an input part containing the structure representation and an output part containing biological activity.

The RDF code (eq 1) of a molecule is a smoothed histogram of all of the intramolecular atom distances that occur and can be interpreted as the probability distribution of finding atom pairs at a distance R.

$$g(R) = \sum_{i=2}^{n}\sum_{j=1}^{i-1} a_i a_j \, \mathrm{e}^{-B(R-r_{ij})^2} \tag{1}$$

where n is the total number of atoms, a_i and a_j are atomic properties of atoms i and j (in our experiments, partial charges q_{tot} were calculated using an in-house protocol based on MPEOE), and r_{ij} is the distance between atoms i and j. B is a smoothing factor which can be interpreted as a temperature parameter defining the fuzziness of atom positions due to thermal movement. In our experiment, this fuzziness is very important because it transforms the atom-distance histograms into smoothed graphs, allowing the RDF codes to be compared using the Euclidian distance measure. We found that a B value of 100 provided reasonable results. The 3D atomic coordinates were calculated using the 3D structure generator CORINA.[10]

As an extension of the initial RDF code concept where the code is calculated between all occurring atom pairs, we calculated the RDF code three times: first, for atom pairs where both atoms have negative charges, second, where one atom has a negative and the other one a positive charge, and third, where both atoms have negative charges. These three codes were concatenated to constitute the final structure descriptor. This was calculated for an R range of between 1.0 and 8.5 Å, which provides a good balance in describing small molecules and not leaving out too much information for larger molecules. For the sake of computation time during network training in the standard mode, we calculated the RDF code in bins of 0.3 Å. Of course, this can be reduced to 0.1 Å or smaller if the analyzed problem requires a higher resolution.

2.2. Artificial Neural Networks. Because the methodology of neural networks has been comprehensively published in journals and textbooks, we will only provide a brief description of the neural network type used in our experiments. CPG neural networks consist of a two-dimensional arrangement of $x \times y$ connected neurons (Figure 1, refs 6 and 7). Each neuron consists of z weights. The number of weights z corresponds to the dimensionality of the molecules' representation containing an input part (structure representation) and an output part (biological activity).

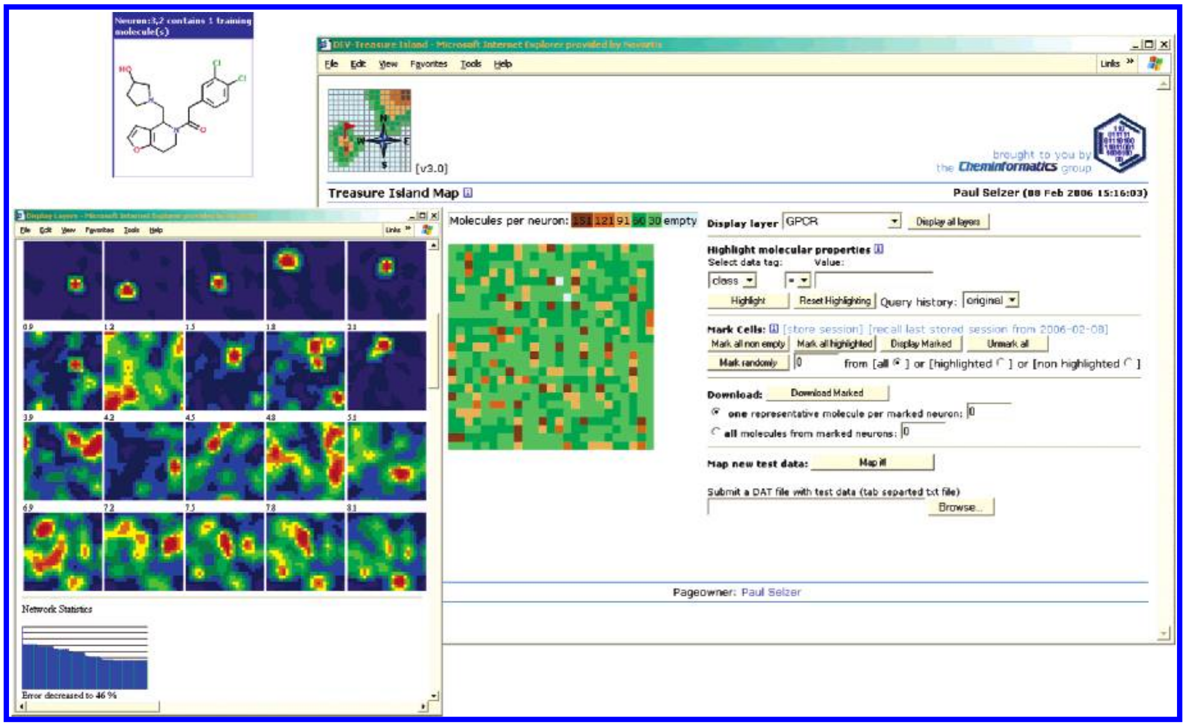
During training, the network learns inductively about the correlation between input and output by analyzing a so-called training set. The training set is presented to the network several times. In each iteration for each molecule, the most similar neuron, the so-called winning neuron, is determined by calculating the Euclidian distance between the structure descriptor and the input parts of the neurons. Then, the neuron weights are adjusted to become more similar to the training data. The winning neuron is adjusted to the highest degree. The neighborhood neurons are adjusted also, whereby the degree of adjustment of a certain neuron decreases with an increasing distance to the winning neuron.

Once trained, the ANN has the ability to predict the property for test set compounds not used during training. In a test run, again the most similar neuron for each test molecule is determined by calculating the Euclidian distance between the neuron weights and the molecular descriptor. Then, the molecular property to be predicted is looked up in the output layer; however, unlike the training process, no weight adjustment is performed.

3. APPLICATIONS AND IMPLEMENTATION OF A WEB TOOL

The use of Web-based cheminformatics tools has a long tradition at Novartis. These tools are integrated into the company's intranet and allow chemists to calculate various types of molecular properties and perform molecule visualization, bioisosteric design, and many other cheminformatics tasks.[11,12] To enhance this tool collection, we have also implemented a Web tool that allows researchers to run neural network experiments and analyze the results interactively (Figure 2). The usage of this tool, which we call "Treasure Island", is very straightforward. It allows the submission of a set of compounds as a SMILES or SD file, or as a generic set of descriptors in text format. Advanced users also have the ability to specify network parameters such as the size, the number of training iterations, and so forth. Otherwise, these parameters are provided automatically.

After the training process, results are displayed as a color-coded map showing the distribution of the compounds among the neurons. An important advantage of this approach is the ability to interactively analyze the results. By moving the


Neuron:3,2 contains 1 training molecule(s)
HO
Cl
Cl
N
N
O
O
DEV-Treasure Island - Microsoft Internet Explorer provided by Novartis
File Edit View Favorites Tools Help
Links
N
W
E
S
[v3.0]
brought to you by the Cheminformatics group
Display Layers - Microsoft Internet Explorer provided by Novartis
0.9
1.2
1.5
1.8
2.1
3.9
4.2
4.5
4.8
5.1
6.9
7.2
7.5
7.8
8.1
Network Statistics
Error decreased to 46 %

Treasure Island Map

Paul Selzer (08 Feb 2006 15:16:03)

Molecules per neuron: 151 121 91 60 30 empty

Display layer GPCR [Display all layers]

Highlight molecular properties

Select data tag: class Value: =

[Highlight] [Reset Highlighting] Query history: original

Mark Cells: [store session] [recall last stored session from 2006-02-08]

[Mark all non empty] [Mark all highlighted] [Display Marked] [Unmark all]

[Mark randomly] 0 from [all ●] or [highlighted ○] or [non highlighted ○]

Download: [Download Marked]

● one representative molecule per marked neuron: 0

○ all molecules from marked neurons: 0

Map new test data: [Map it!]

Submit a DAT file with test data (tab separted txt file)

[Browse...]

Pageowner: Paul Selzer

Figure 2. Output page of the neural network Web tool, which features the display of chemical structures and corresponding properties and the download of cluster results.

mouse over the neuron cells, the centroid compound is displayed. By clicking on the neurons, they can be marked and all data associated with these neurons are displayed. It is also possible to download structures and data from certain network regions, for example, neurons containing molecules with high biological activity. The comparison of input and output layers allows an evaluation of the importance of certain input variables with regard to the output. This information and the analysis of the network error during training (graph in the lower left-hand corner of Figure 2) can be used to improve the quality of the model by discarding input variables that do not appear to be correlated with the output property. Last, the results can be archived on the server, documented, and shared with other scientists.

In the following sections, we summarize several examples of where the Web-based neural network methodology has been successfully applied.

3.1. Prediction of Biological Activity. As described in a previous publication,[13] the method was successfully used to predict the G-protein-coupled receptor (GPCR)-ligand likeness of compounds by training a network with 1709 known GPCR ligands and 24 870 common druglike molecules.

In addition to the structure descriptor (described in section 2.1), a binary classifier was appended to describe whether molecules were GPCR substrates or not. However, the only information used for the training was the structure descriptor. The GPCR class information only was used for analyzing the clustering and prediction results. Figure 3 shows the weights from the output layer (binary GPCR classifier). We can observe a very good clustering of the GPCR ligands with respect to the inactive compounds. Furthermore, it is possible to efficiently separate peptidic ligands (upper left-hand corner) from aminergic compounds (center).

For the prediction process, test set compounds were mapped onto the trained network and, for each of these, the most similar neuron was determined. The weight of the GPCR layer of this neuron constitutes the GPCR-ligand-likeness score of the respective test set compound. Using this approach, we were able to predict 71% of the active GPCR compounds correctly from a data set that contained only 5.9% of the active compounds in total.

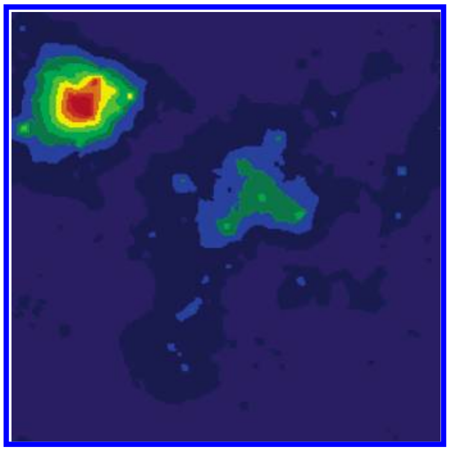


Figure 3. Result map, which shows a clear separation between GPCR ligands and inactive compounds. Peptidic ligands are clustered in the upper left-hand corner, aminergic ligands in the center of the map.

3.2. Compound Selection for Screening. A very prominent task in pharmaceutical research is the selection of compounds for screening. Pharmaceutical companies are regularly buying large numbers of compounds from various commercial providers to extend their in-house collections and to feed high-throughput screening robots. These compounds are selected on the basis of various criteria—novelty relative to the company collection, calculated properties, and various bioactivity models. ANNs can support this process in two ways.

In the first case, the network is trained with a set of compounds with known activity. The set has to contain active and inactive compounds so the network can develop a model to discriminate between these two groups. Then, a set of compounds with unknown activities can be mapped into the trained network. Those compounds being assigned to neurons containing active compounds from the training set have a high probability of showing a similar activity and should therefore be suggested for screening. This approach can be applied if a number of compounds have already been screened in a particular assay and the selection of compounds of unknown activity (e.g., from the company's compound collection) are required for further screening.

A second approach can be used to select promising screening candidates from a set of compounds with unknown activities, for example, a large compound collection from a vendor catalog. This compound collection is used during training to create the initial map. A set of compounds (or a single compound) with known (high) activity, for example, from a publication or a patent, is mapped into the trained network. Compounds from the initial data collection that are located on neurons to which the active compounds were assigned have a high probability of showing a similar activity and are, therefore, promising candidates to purchase.

For both approaches, the quality of the results increases with the similarity between the compounds in the (initial) training data set and the compounds in the mapping data set.

Both approaches are complementary, and the appropriate approach has to be chosen on the basis of the task and the data situation.

3.3. Selection of a Representative Subset. The neural network methodology can also be used to select a representative subset from a large set of compounds (e.g., a collection of several combinatorial libraries). To perform such an experiment, the neural network was trained with the entire compound collection. The number of neurons should be set equal to the number of compounds that are required for selection. If the number of molecules in the initial compound collection is much larger than the number of neurons, usually all neurons should be occupied by at least one compound. If the initial set is not too large and exhibits only a low diversity, then the number of neurons required will be higher than the number of compounds to be selected because many empty neurons (i.e., neurons to which no molecule has been assigned during training) will be present. After training, the centroid compound for each neuron is identified as being the compound having the lowest Euclidian distance between its structure descriptor and the neuron weights. The resultant collection of such centroid molecules constitutes the representative subset.

We utilized this technique to select a subset of 5000 compounds from a collection of combinatorial libraries containing around 100 000 compounds in total. The benefit of this approach may be seen in Figure 4, which shows how many compounds belong to libraries, respectively selected subsets, of certain diversity ranges. The similarity measure applied was the Tanimoto coefficient of the molecular fingerprints. The light gray bars show the diversity distribution of the initial library collection; the dark gray bars show the distribution of the randomly selected subset, and the black bars show the distribution of the subset selected with the neural network. Figure 4 shows the distributions for sets of libraries of low, medium, and high diversity. Additionally, the distribution of one small and highly diverse library is displayed. We observe that the random selection provides a distribution very similar to the original data set, whereas the neural network selection picks less compounds from low and medium diversity libraries and more compounds from highly diverse libraries. These findings are also supported by the distribution for one small but highly diverse library where we observe that significantly more compounds were picked by the neural network approach than by random selection.

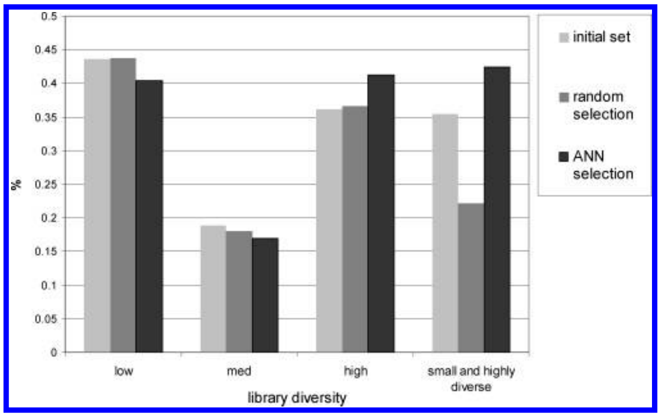


Figure 4. Chart showing the result of a subset selection. It can be observed that the the neural network technique favors compounds from diverse libraries, even though most of the compounds from the initial data set belong to libraries with less diversity.

This indicates that the neural network approach ensures that the selected subset covers a chemical space that is representative of the initial large collection of compounds. In the case of random selection, there is an increased probability that compounds from highly diverse but small libraries would be under-represented.

4. CONCLUSIONS

Artificial neural networks provide a wide range of potential applications in pharmaceutical research. Typical applications are the prediction of biological properties, the selection of screening candidates, and the generation of a representative subset from a large compound collection such as a combinatorial library.

Obviously, neural networks have to compete with other statistical methods such as partial-least-squares analysis, support vector machines, and nearest neighbor methods, which might be faster while performing equally well or, in some cases, even better. However, the big advantage of Kohonen and CPG neural networks is that they provide quick and intuitive feedback about the results of the cheminformatics experiment, for example, about the quality of the structure descriptors, the correlation of input and output, or the contribution of a particular input property to the output. This visual feedback is an important factor for the acceptance of this method because it meets the needs of the researcher, who is often trained to process information, like chemical structures, in a graphical manner.

ACKNOWLEDGMENT

We would like to acknowledge Johnny Gasteiger for raising our interest in the application of neural networks in chemistry. P.S. especially thanks Johnny for supervising his Ph.D. studies at the Computer-Chemistry-Center in Erlangen.

Furthermore, we thank Philipp Floersheim for his input on the improvement of the RDF code and Bernhard Rohde, Jörg Mühlbacher, Stephen Jelfs, Ansgar Schuffenhauer, Hans-Jörg Roth, and other colleagues at Novartis for many fruitful discussions and the collaboration on the development of our cheminformatics Web-tool portfolio.

REFERENCES AND NOTES

(1) Gasteiger, J.; Engel, T. *Chemoinformatics*; Wiley-VCH: Weinheim, Germany, 2003.

(2) Gasteiger, J. *Handbook of Chemoinformatics*; Wiley-VCH: Weinheim, Germany, 2003.

(3) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Kluwer Academic Publishers: Dodrecht, The Netherlands, 2003.

(4) Selzer, P.; Rohde, B.; Ertl, P. Cheminformatik und Data Warehousing: Forschen mit dem Intranet. *Nachr. Chem., Tech. Lab.* **2000**, *48*, 1471−1475.

(5) Kohonen, T. *Self-Organizing and Associative Memory*, 3rd ed.; Springer-Verlag: Berlin, 1989.

(6) Hecht-Nielsen, R. Counterpropagation Networks. *Appl. Opt.* **1987**, *26*, 4979−4984.

(7) Zupan; J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinheim, Germany, 1999.

(8) Steinhauer, V.; Gasteiger, J. Obtaining the 3D Structure from Infrared Spectra of Organic Compounds Using Neural Networks. In *Software-Entwicklung in der Chemie 11*; Fels, G., Schubert, V., Eds.; Gesellschaft Deutscher Chemiker: Frankfurt/Main, Germany, 1997.

(9) Gasteiger, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. Finding the 3D Structure of a Molecule in Its IR Spectrum. *Fresenius' J. Anal. Chem.* **1997**, *359*, 50−55.

(10) 3D Structure Generator CORINA. www.mol-net.de (accessed Feb 2006).

(11) Ertl, P.; Mühlbacher, J.; Rohde, B.; Selzer, P. Web-based Cheminformatics and Molecular Property Prediction Tools supporting Drug Design and Development at Novartis. *SAR QSAR Environ. Res.* **2003**, *14*, 321−328.

(12) Ertl, P.; Selzer, P.; Mühlbacher, J. Web-based cheminformatics tools deployed via corporate Intranets. *Drug Discovery Today: BIOSILICO.* **2004**, *2*, 201−207.

(13) Selzer, P.; Ertl, P. Identification and Classification of GPCR Ligands Using Self-Organizing Neural Networks. *QSAR Comb. Sci.* **2005**, *24*, 270−276.

CI0600657