Proceedings of the 5th WSEAS/IASME Int. Conf. on SYSTEMS THEORY and SCIENTIFIC COMPUTATION, Malta, September 15-17, 2005 (pp12-16)

A Classification Scheme for the Prediction of Essential Chemical and Biological Properties based on the Classical Neural Network Approach

MARKUS BORSCHBACH
Dept. of Mathematics and Natural Science, Institute for Computer Science,
University of Münster, Einsteinstr. 62, D-48149 Münster, GERMANY
http://cs.uni-muenster.de/u/flyer/

Abstract: - Prediction of biological and chemical features of a given chemical structure is a challenging problem for the existing nonlinear mapping performed by neural networks. In combinatorial chemistry, computational approaches are capable to significantly decrease the necessary amounts of synthesis for the development of a specific chemical or biological drug. Therefore, the main goal is to distinguish appropriate descriptors from insignificant ones. The experimental design for the classical nonlinear neural network mapping for the approximation of five descriptors and the corresponding reaction of the immune system for the drug development are reported briefly. The results for the different descriptors are presented in comparison.

Key-Words: Drug Design, Neural Network, Computational Approach for Combinatorial Chemistry.

1 Introduction

Since the last decade, the interest in computational approaches supporting the design and development of drugs is growing steadily. Besides the well established methods, the development cycles of drugs give the opportunity of challenging problems for well known AI (Artificial Intelligence)-concepts, e.g. see [1], [8], [11]. From a chemical hypothesis point of view, exists for almost every type of biological activity an appropriate molecule with a specified chemical structure. The medical aim of drug discovery is to identify a bio-active molecular structure that systematically interfere pathological processes in a positively manner, cure a disease or prevent the initial conditions of a disease. For both purposes, the idea of AI-approaches is to improve the development process by reducing the enormous search space for a significant biological activity for a specified action (for instance, see [13]).

The emulation of development steps for combinatorial synthesis has attracted a lot of research in combinatorial chemistry, e.g. see [21] for an overview. The goal is an iterative generation of smaller sample sets in combination with an experimentally determined biological response. In comparison to the classical approaches of enormous synthetic libraries, the evolutionary process gives the opportunity of preventing pure combinatorial generation and evaluation of instances for a given structure based algebraic molecule representation. Especially, if the combinatorial approach becomes physically unfeasible for a growing magnitude of the underlying chemical structure, e.g. the length of the peptid or the number of amino acids for every peptid.
The experimentally determined biological response for each succeeding generation is used for the fitness evaluation and selection of the predecessors according to the applied evolutionary operators. The main achievement of the evolutionary process itself is the internal coding of the chemical structure and its dependency on the generation of either chemical as similar as possible molecular structures or mostly divers compounds. As a further step on minimization of the necessary number of synthesized experimental responses for instances of a chemical structure, a certain property is predicted by a trained neural network. The verification of this prediction leads to a comparison of different descriptors [4]. Selection and development of descriptors for a specific biological activity have been identified before as the prerequisite for the next revolution in drug design [10]. The classical neural network architecture, a backpropagation neural network (BNN, see [2], [22], [10], [4] for configuration issues), is used for the prediction of immune assay responses. This paper is focusing on the comparison of different descriptors for the prediction with neural networks and develops a concept for a more sophisticated classification scheme. The organization of this paper is as follows. Section 2 reveals the experimental details and the medical motivation for the neural network prediction. Accordingly, section 3 compares the prediction quality based on neural networks approximating different descriptor and two independent effector sides. Encouraged by the descriptor selection, an advanced classification scheme is presented in section 4.

2 Experimental Design Selection

Two general distinguished applications cases for the nonlinear mapping of neural networks in the context of physical-chemical descriptors does exist. The goal of both cases is to approximate a highly nonlinear mapping among an input variable I and the output variable P. For both cases, the predicted output p is denoted as a function of real number values v for any value i of the input variable $\{p(i) \mid v \in I\!R\}$. The input variable I at an instance t represents an instance $i(t)$ of the predefined chemical structure. Examples of the estimated experimental responses incorporate almost any subject of QSAR (quantitative structure-activity relationship) and QPAR (quantitative property activity relationship), or any other experimental determined response, e.g. the prediction of antibody immune response [5] to name only a specific one, focused in some experimental results later.

Based on a given pool of relevant descriptors [15], [14], [3], [17], the main goal is to determine a selection of descriptors by the importance of the approximated prediction. The straight forward approach for determination of the significance of a descriptor, is to estimate the performance of a single neural network for each descriptor. Therefore, any instance of the predefined chemical structure is coded by the chosen descriptor for a specific neural network.

The details and results of the comparison of neural networks for prediction based on values for a given estimation goal can be found in the following section (3).

An alternative approach for estimating a given goal is to use a subset of descriptors to code each instance $i(t)$ of the input variable I, each descriptor coding a component of the input vector $\vec{i}(t)$. The main destination of this approach is to determine the most significant subset of descriptors [9], [20]. In Section 4, a concept for compromising both approaches is presented.

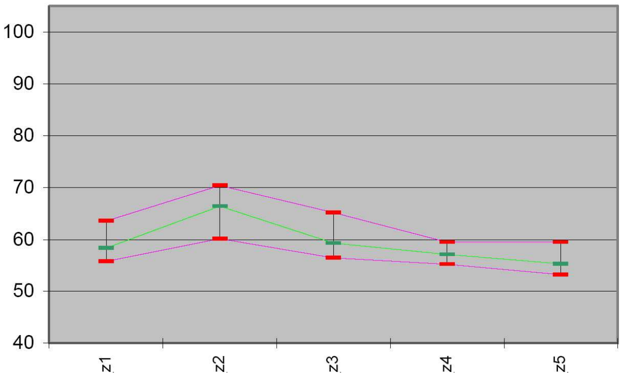

Fig. 1: Different Immune response prediction qualities depending on a neural network for each descriptor $Z_1 - Z_5$

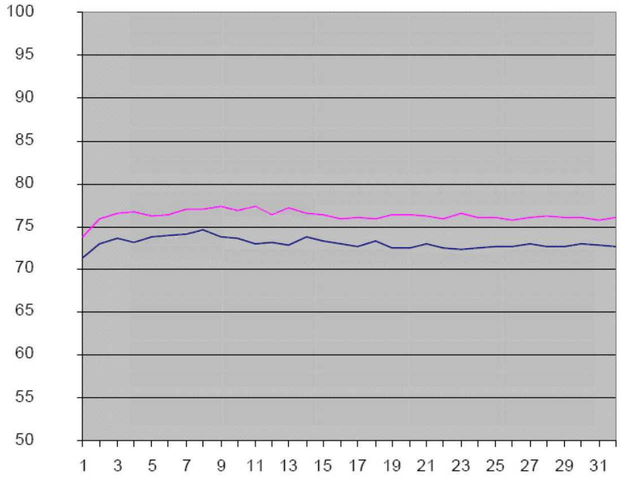

Fig. 2: Comparison of immune response predictions in blood (black) and dejection (grey) based on an average of all approximated descriptors $Z_1 - Z_5$

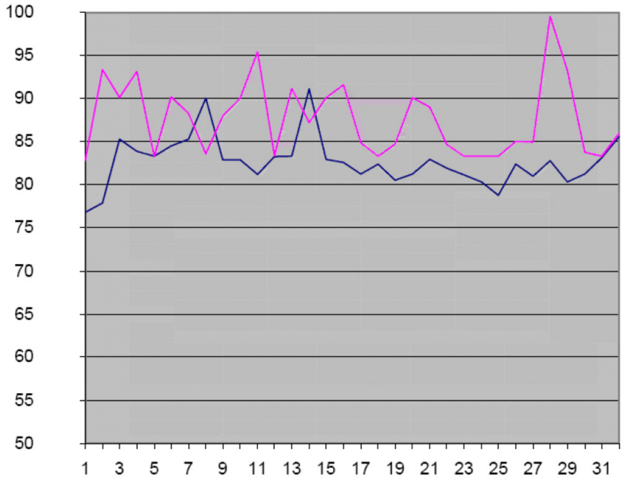

Fig. 3: Comparison of immune response predictions based on the maximum quality of a (not necessary the same) descriptor approximation for each value

3 Comparison of Descriptors

The estimated experimental responses were encouraged by data sets for mucosal antibody immune responses for intragastic immunization [5]. Accordingly, the input variable I for the neural network is coded by different sequences of amino acids. A single neural network is trained for each of the preselected descriptors [12], [18]. For each element of the chemical description space, the components of a instance vector $\vec{i}(t)$ of the input variable is determined by the descriptor value for the given amino acid. For example, if the first amino acid in a given sequence is Alanin, the input value for the first neuron of the input layer (first component of vector $\vec{i}(t)$) depends on the descriptor selection: for $Z1 : 0.24$ and for $Z2 : -2, 32$. Therefore, based on the descriptor each sequence represents a different input vector property for the neural network.

Accordingly, the challenging question can be summarized as follows: Which neural network approximation based on a specific descriptor is estimating the immune response as close as possible for a special effector side ? For the biological and medicinal details and motivation see [5]. In figure 1, the achieved prediction quality for the immune response of each descriptor is illustrated. The values for the minimum, the average and the maximum quality for each descriptor are interconnected by a line.

For each approximated descriptor, an optimized feed-forward neural network structure was determined by simulation of resilent backpropagation, refer to [2], [22], [10], [4] for configuration issues. The results strictly emphasize the use of descriptor $Z2$.
In the following, the quality of the prediction is calculated as the average among the prediction of a single network for each descriptor. The first intention, is to compare different underlying confidence levels of different data sets [7].

The second intention, is to compare the prediction quality depending on the difference of the immune system response, detected either in the blood or the dejection. Figure 2 compares the prediction on single complete test data set for both. The black graph represents the prediction quality values for the responses in the blood and the grey graph the corresponding predictions for the dejection. In comparison, the prediction quality for blood is strictly ahead of the values for the dejection.
Compared to data sets for a confidence level of 99% and another data set of $97,5\%$ (cut-off), the illustrated prediction quality based on the average and the maximum is the most superior one. For the issue of cut-off comparison refer to [16]. In addition, figure 3 summarizes the maximum quality among all networks, divided in a single graph for blood and dejection.

Moreover, figure 3 is used to compare different cut-off-versions with a corresponding confidence factor. The main question arises by a comparison of figure 3 and figure 2 : Is there any improvement for the prediction quality if the maximum quality of more than one descriptor can be incorporated ? Therefore, a concept and results for selecting and combining the superior networks are discussed in the following section.

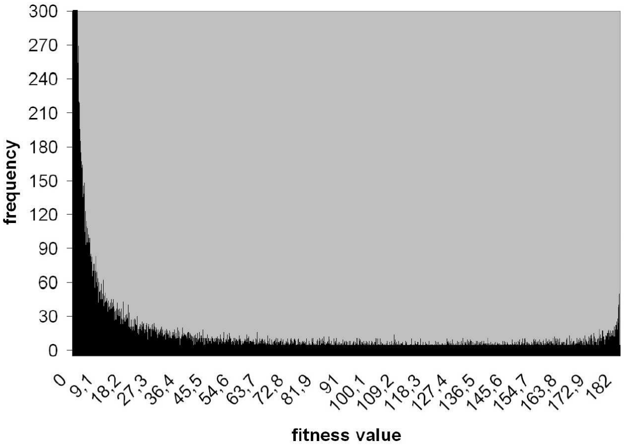

Fig. 4: Distribution of 50000 randomly sequences

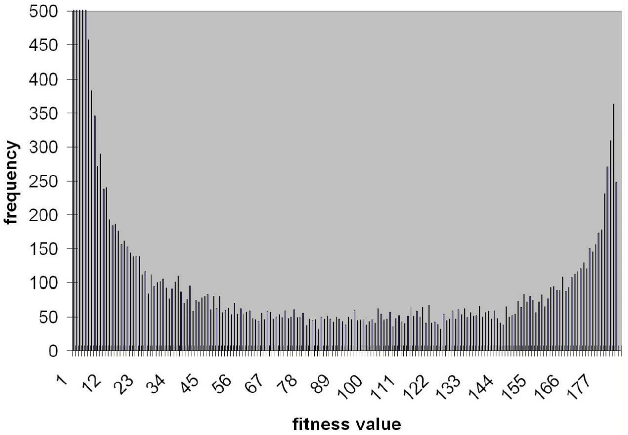

Fig. 5: Distribution of 32768 randomly sequences, restricted to two a.-acids at each position

4 Classification Scheme

The comparison scheme based on the maximum, average and minimum prediction quality of different descriptors is accomplished by a systematic testing of responses to randomly distributed values of the input space $i(t)$. For instance, the sequences of peptids for the immune response prediction example consists of 15 amino acids. If the number of amino acids for each position is 20 different amino acids, the number of different input vectors $\vec{i}(t)$ coded by a predefined descriptor is 20^{15}.

A set of 50.000 randomly distributed input vectors is generated to evaluate the prediction quality of a given network approximation. Moreover, the network response has been traced. In figure 4, the response of a neural network approximating a continuous experimental activity is visualized for every vector of the generated input space. Despite a high prediction quality for the test data set, the prediction quality has been further evaluated by 32768 randomly generated sequences of amino acid restricted to different subsets of only two amino acids at each sequence position.
The main conclusion from subsets of single amino acids with a high number of occurrences in sequences with a high experimental activity is visible in figure 5. The shape of the distribution of values is almost equivalent to the one based on all of the twenty amino acid. Of course, this kind of shape was restricted to subsets of amino acids with a high occurrence in sequences a corresponding high experimen-

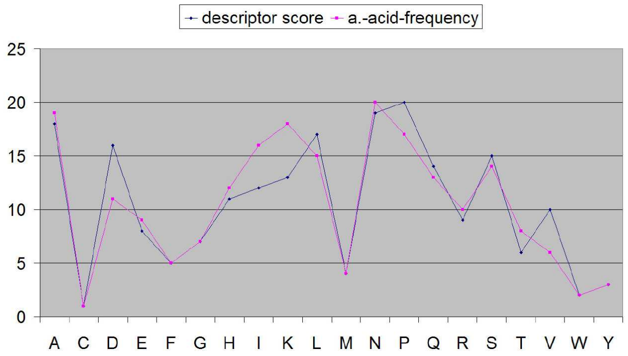

Fig. 6: Correlation of descriptor value and occurrence for target sequences

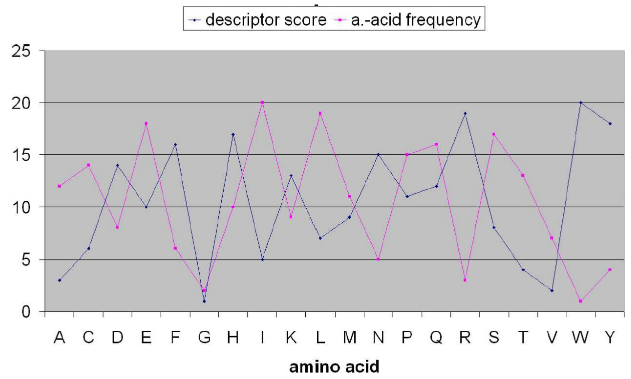

Fig. 7: Partly uncorrelated descriptor value and occurrence for target sequences

tal activity. For a further analysis of these subsets of amino acids, the descriptor value and the occurrence of a amino acid for sequences with a high experimental activity have been evaluated. Figure 6 illustrates the significant correlation among the descriptor value and the occurrence of these amino acids for sequences with high experimental activity.

An extended evaluation of the network prediction quality leads to an inevitable conclusion. Despite the achieved prediction quality on early test sets, the network prediction quality is rather weak.

Based on an extended training and a larger test set, a more significant network approximation leads to a more sophisticated prediction quality. Therefore, figure 7 visualized a partly uncorrelated significance of the descriptor value and the corresponding number of occurrences.

5 Conclusion

Like all ideas of major influence on the development of science, the idea of a rigorous test scheme is common sense in many different areas. For the prediction quality based on neural network approximation of descriptors, two different schemes have been proposed. While the first is capable of thoroughly comparing prediction quality among different descriptors, the second gives the opportunity for immanent prediction quality tests. In comparison to the related work based on SOM [19], both test schemes give the opportunity to maintain the most significant neural network predictions. Strongly encouraged is a performance contest among SOM&CPNN prediction and a subset of classical neural networks selected by the presented schemes. Both issues are topic of ongoing and future research.

References

[1] A.C. Anderson and D. Wright, The Design and docking of Virtual Compound Librariers to Structures of Drug Targets, *Current Computer-Aided Drug Design I*, 2005, pp. 103–127.

[2] J. Devillers, Strengths and Weakness of the Backpropagation Neural Network in QSAR and QSPR Studies, In *Neural Networks in QSAR and Drug Design*, London: Academic Press, 1996.

[3] M.V. Diudea, *QSPR/QSAR Studies by Molecular Descriptors*, NY: Nova Science Publishers, 2001.

[4] L. Douali, D. Villemin, and D.Cherqaoui, Exploring QSAR of Non-Nucleoside Reverse Transcriptase Inhibitors by Neural Networks: TIBO Derivatives, *Int. J. Mol. Sci.*, vol. 5, 2004, pp. 48–55.

[5] D. Externest, B. Meckelein, M.A. Schmidt and A. Frey: Correlations between Antibody Immune Responses at Different Mucosal Effector Sites Are Controlled by Antigen Type and Dosage, *Infect. Immun.* vol. 68, 2000, pp. 3830–3839.

[6] D. Externest, *Analysis of immun responses* (in German: ”Analyse der Immunantworten und Erkennung linearer B-Zell Epitope nach mukosaler und systemischer Immunisierung: Neue Erkenntnisse für die Entwicklung von mukosalen Peptidvakzinen”, Münster: PhD-thesis, 2000.

[7] A. Frey, J.D. Canzio and D. Zurakowski, A statistically defined endpoint titer determination method for immunoassays, *J. of Immunological Methods* 221, 1998, pp. 35–41.

[8] E.M. Krovat, T. Steindl and T. Langer, Recent Advances in Docking and Scoring, *Current Computer-Aided Drug Design I*, 2005, pp. 93–102, 2005.

[9] R. Leardi, *Nature-inspired methods in chemometrics: Genetic algorithms and artificial neural networks*, Amsterdam: Elsevier, 2003

[10] D.T. Manallack and D.J. Livingstone, Neural networks in drug discovery: have they lived up to their promise ?, *Eur. J. Med. Chem.*, vol. 34, pp. 195–208, 1999.

[11] A. Odugauwa, A. Tiwari and R. Roy, Overview of Soft Computing Techniques in Lead Identification and Optimization for the Drug Discovery Proces. *Proc. of Workshop on challenges in real world optimisation using evolutionary*, Cranfield University: Decision Engineering Report Series,, R. Roy and C. Kerr (eds.), pp. 7–16, 2004.

[12] M. Sandberg, New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids, *J. Med. Chem.*, vol. 41, pp. 2481–2491, 1999.

[13] G. Schneider and S.-S. So, *Adaptive systems in drug design*, Georgetown, Tex. : Eurekah.com/Landes Bioscience, (Biotechnology intelligence unit ; 5) ISBN: 1-587-06118-X, 2003.

[14] G. Schneider and S.-S. So, *Modeling Structure-Activity Relationship*, In ”Adaptive Systems in Drug Design”, G. Schneider and S.-S. So (eds.), 2003.

[15] G. Schneider and S.-S. So, *Analysis of Chemical Space*. In ”Adaptive Systems in Drug Design”, G. Schneider and S.-S. So (eds.), 2003.

[16] W. Schomburg, *Pattern recognition for the prediction of immune responses based on neural networks* (in German). Münster University: Dissertation submitted for a Diploma, 2002.

[17] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, vol.11, 2000.

[18] M. S. Venkatarajan and W. Braun, *New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties*, Springer-Verlag, 2001.

[19] M. Vracko, Kohonen Artificial Neural Network and Counter Propagation Neural Network in Molecular Structure-Toxicity Studies, *Current Computer-Aided Drug Design I*, pp. 73–78, 2005.

[20] R. Wehrens and L.M.C. Buydens, Chemometrics, In *Evolutionary Algorithms in Molecular Design*, Clark, D.E. (eds.), 2000.

[21] L. Weber, Evolutionary Combinatorial Chemistry: Application of Genetic Algorithms, In *DDT*, vol. 3, pp. 379–385, 1998.

[22] J.H. Wikel, E.R. Dow, and M. Heathman, *Interpretative neural networks for QSAR*. Network Sci. [Electronic Publication], (URL: http://www.awod.com/netsci/Issues/Mar96/feature1.html), No pp. Given, 1996.