WCCI 2012 IEEE World Congress on Computational Intelligence
June, 10-15, 2012 - Brisbane, Australia IJCNN

An Improved Neural Network Ensemble Model of Aldose Reductase Inhibitory Activity

B. Keshavarz Hedayati, R. Parra-Hernandez,
E. M. Laxdal, N. J. Dimopoulos
Department of Electrical and Computer Engineering
University of Victoria, B.C., Canada
E-mail:{babak,rafaelph,elaxdal,nikitas}@ece.uvic.ca

P. Alexiou, V. J. Demopoulos
Department of Pharmaceutical Chemistry
Aristotle University of Thessaloniki, Greece
E-mail: {ksenia, vdem}@pharm.auth.gr

***Abstract*— In this paper, we improve the results based on a Neural Network-based model that predicts an enzyme (Aldose Reductase) inhibitory activity of a group of compounds. The improvement is due to the judicial selection of ensembles of trained Neural Networks to contribute to the final model. The method is validated on a family of compounds that is different from the families which were used in the training of the model. The results confirm an accurate, chemical-family-independent method that can predict Aldose Reductase inhibitory activity with excellent accuracy.**

I. Introduction

This present paper is a continuation of the work described in [15] and [16]. There, we reported the results in the construction of a NN-based prediction model for an enzyme inhibitory activity of a number of chemical compounds and its validation using another set of unknown compounds. The focus of the present work is to refine and expand the methods presented earlier leading to an improved prediction of Aldose Reductase Inhibitory (ARI) Activity. We shall validate the results obtained using an enhanced validation set of actual experimentally derived activity data in a blind experiment. We shall show that the new refined model predicts the (Aldose Reductase) Inhibitory Activity exceptionally well.

A key step in the construction of a NN-based prediction model is the training of the NNs. Our training mechanism and its applications reported in [10], [15], share the fact that the information (data or exemplars) used came from observations on real systems, the amount of information at hand was sparse, and the relationships modelled were complex.

Here, we focus in modelling pharmacological activity based on structural properties of a chemical compound. Specifically, we are interested in modelling the Aldose Reductase (AR) enzyme inhibitory activity. The inhibition of the AR enzyme is considered to be an approach to control diabetic complications, ischemia, abnormal vascular smooth cell proliferation, cancers, and mood disorders [13], [2] [11].

A large body of literature reports on the modelling of possible relationships between biological/chemical/pharmacological activity and the structure of a compound [3], [17]. The method is based on deriving a model of the activity under study based on descriptors of the compound in question. The descriptors can be classified as *topological*, *geometric*, *electronic*, *physicochemical*, or *hybrid*. The activity may be biological, its value being obtained as an assay on specific biological target(s), or it may be associated to specific properties such as chromatographic ones. The general method is known as quantitative structure-activity relationship (QSAR). Linear and non-linear models, including NNs [14] are used.

The main difficulty in research utilizing experimentally derived sets of exemplars is the small number of exemplars which makes training and generalization difficult. We shall show, that our methods are very successful in such problems.

This work is presented as follows. In Section II, we briefly describe the available data. In Section III we present an overview of the method we have developed. In Sections IV, V, and VI, we describe the methods used to select the training sets, train the NNs and establish the ensembles of trained NNs that will eventually yield the activity model. In Section VII we present the method of selecting the ensembles of trained NNs that will contribute to the final model. The model is evaluated on a blind set of compounds and the results are reported in Section VIII. We conclude with section IX.

II. Aldose Reductase Inhibitors Data

Generally, in the construction of structure-activity relationship models, one uses data from a single family of compounds. However, in this work we used data from three different families of chemical compounds. This is an important difference in that our method leads to a general model encompassing several families of compounds. The families we used are: nitrophenyl derivatives, with 19 compounds [4]; phenolic derivatives, with 27 compounds [5]; and pyridazine derivatives, with 15 compounds [12] for a total of 61 compounds.

These 61 compounds were used to develop our models. The results were included in [15]. Subsequently, the descriptors of a further 19 unknown compounds were made available, and these were used for the blind validation tests as we shall discuss in Section VIII. We shall refer to these 19 unknown compounds as the *Blind Validation Set*

Each compound is described by a set of 17 molecular descriptors: Surface Area and Volume of Electron Density, Electrostatic Potentials (ESPmax, ESPmin and ESPdiff), Energy parameters (EHOMO, ELUMO, Ebandgap), Dipole Moment,

Hydration Energy, Water Accessible Surface Area, Water Accesible Volume, Polar Molecular Surface Area, Lipophilicity, Molar Refractivity, Polarizability and the number of hydrogen bond acceptor sites. The AR inhibitory activity (output data) is determined as the concentration exhibited at a 50% inhibition (IC_{50}) of the enzyme isolated from rat lenses, covered a wide range between 10^{-4}M and 10^{-9}M and was converted to pIC_{50} ($= -log(IC_{50})$)values.

Of the 61 compounds, four compounds were randomly selected to form a *validation set*. The remaining 57 compounds were further divided into two sets: a *Training Set* and a *Test Set* which were constructed using *guided selection* [10].

The descriptors were processed with MATLAB's NN Toolbox[1] function *premnmx.m* to normalize the inputs in [-1,1].

The resulting normalization transformation vectors were also used to transform the *validation set* and the *Blind Validation Set*. No preprocessing steps were performed on the output data vector, that is, on the AR inhibitory activity values.

No further pre-processing that could have eliminated parameters with very small variability was employed, since it was deemed that the dimensionality of the input space (17) was small enough for the NN to handle computationally.

III. Summary of the Method

Since our method involves several steps, we would like to summarize it at this point so it would be easier for the reader to follow.

The aim of our method is to devise a model that would be able to model the inhibitory activity of Aldose Reductase and generalize so that it could predict the Aldose Reductase Inhibitory Activity of unknown compounds.

We develop our model based on the activity of a set of known compounds (as discussed in Section II, this set includes 61 compounds). As a first step, and in order to be able to determine the generalization abilities of our model, we randomly partition this set of compounds into a *validation set* (comprising 4 compounds) and the *residual set* comprising 57 compounds.

The *residual set* is now used to develop the model. To each of the compounds there corresponds an exemplar comprising a vector of descriptors and the expected (experimentally derived) Aldose Reductase Inhibitory Activity. Given that the set of exemplars is quite sparse, we employ the *guided selection mechanism* we developed previously [10] to select the *Training* and *Test* sets. A set of neural networks are now trained. Subsequently, we thin this set of neural networks keeping only the ones that show good generalization abilities. This process employs the *Sensitivity Heuristic* we have developed, and utilizes the *Test* and *Validation* sets.

The resulting set of trained neural networks, corresponding to a particular partition of the original set of compounds, is called a *Sequence*. We repeat the partition process several times and each resulting *Sequence* encompasses a different "view" of the eventual model. The question now is to devise a method of combining these *Sequences* to a coherent general model of the Aldose Reductase Inhibitory Activity.

The general idea is to ensure that collectively the responses of the trained neural networks of all the *Sequences* behave somewhat similarly. *Sequences* whose responses are heavily weighted to the extremes of the response intervals, or deviate from the norm, are excluded or their influence to the final result is diminished. We shall present the method of determining the influence and participation of individual *Sequences* in Section VII.

We shall also present the particulars of the *Sensitivity Heuristic* in Section V

IV. NN Type and Training Mechanism

We selected a NN with six neurons in the hidden layer, an activation function defined as $y = 2/(1 + exp(-2n)) - 1$ and a single output with a linear activation function. The number of neurones in the hidden layer was selected after experimentation to ensure proper learning and generalization [10].

The *guided selection mechanism* [10] selectively divides the set of exemplars into two subsets namely the *Training* and the *Test Sets*. This selection mechanism, although similar, should not be confused with methods for estimating generalization errors, including cross-validation [8]. Our *guided selection mechanism* is used to iteratively construct a training subset from the available sparse set of exemplars, which will result in generalization-capable Neural Networks.

V. NN-based models post-processing

In this study, the guided mechanism generated 350 candidate *Training* and *Test Sets* pairs. Associated with each pair, our method also generates a score s_i[2]. This score is used by our method to guide our algorithm and it is a measure of the generalization capability of the NNs trained and evaluated on the corresponding pair of sets of exemplars. We select the "best" pair of *Training* and *Test Sets* by choosing the pair that has the largest *Test Set* and the highest score.

For the chosen pair of sets of exemplars we retrieved the set of NNs our algorithm generated. There were 500 NNs associated with each pair of exemplar sets. From this set of trained NNs, we are interested in choosing one (or more) NNs that will generalize well and construct an ensemble to be used in predicting ARI; similar ensemble construction strategies have been used previously [6].

[1]MATLAB is a registered trademark of The Math Works,Inc.

[2]For each *Training Set* i, we train $q = 500$ NNs, starting at different initial conditions. For each set of exemplars, we calculate a success *score* s_i as:

$$s_i = \frac{1}{q} \sum_{j=1}^{q} h_i(j) \tag{1}$$

where:

$$h_i(j) = \begin{cases} 1 & \text{if} \quad e_{i,j} \leq 0.06 \\ 0 & \text{if} \quad e_{i,j} > 0.06 \end{cases} \tag{2}$$

$e_{i,j}$ is the normalized error obtained when the NN obtained in training session j, using the *Training Set* i, was tested on the corresponding *Test Set* i.

We have already selected four exemplars which were excluded from the set of exemplars used for our guided selection method and which constitute the *validation set*. We will proceed now and examine the set of the trained NNs as to their ability to generalize based on the *validation set*.

Given that the exemplars in the *validation set* have not be "seen" by the NNs, there are NNs which will have responses that will are "abnormal". The *Sensitivity Heuristic* [15] described below, aims to eliminate those NNs that have such "abnormal" responses. The method perturbs the inputs to the NNs, and then discards NNs that have responses that are statistically outside the "norm", retaining NNs with self consistent responses. Note that this method relies only on the input/output transformation as implemented by the trained NN and does not consider the expected responses.

Sensitivity Heuristic:

1) Perturb the exemplars in the *Training Set* by adding uniformly distributed noise to the values of the descriptors and compute the resulting NN response. That is, create a *noise-Training Set* made of noise-compounds and compute the resulting NN response.
2) For each noise-compound, compute the *average-per-noise-compound response* as follows. Eliminate, in a per noise-compound basis, only the corresponding NNs response to that particular noise-compound that are *non-consistent*; the *average-per-noise-compound response* is the average response of the remaining responses (i.e., the skewness of the remaining response distribution is close to zero). When computing an *average-per-noise-compound response*, only the corresponding noise-compound response of some NNs is eliminated, the response to other noise-compounds remains available. For each m noise-compound, name the *average-per-noise-compound response* as $ANR(m)$.
3) For each NN j as of step 1 above, compute the neural network's average error for all noise-compounds in the noise-Training set. That is

$$\text{mean error}_j = \left(\sum_{m=1}^{M} \text{error}_j(m) \right) / M, \qquad (3)$$

where $\text{error}_j(m) = |ANR(m) - NN_j(m)|$
4) Now, eliminate the NNs so that the skewness of the average-error is close to zero.
5) Considering the selected NNs from the previous step as the new total number of NNs, repeat steps $1-4$ but now using the *validation set* instead of the *Training Set*.
6) Construct the NN-based predictor as the ensemble of the remaining NNs from step 5.

The *Sensitivity Heuristic* was instantiated as follows. The added uniformly distributed noise had a mean value of zero. In a first stage, Steps $1-4$ were applied several times, each time the new total number of NNs decreases (Step 4). Each time the amount of noise added was decreased. The first time, the noise uniform distribution was in the interval of (-0.2,0.2), then the interval was reduced by 90% each time. Steps $1-4$ were applied as many times as needed to obtain a new total number of NNs equal to 100 (i.e., 20% of the original 500 NNs). Next, in a second stage, we applied Step 5 several times. As in the first stage, each time the new total number of NNs decreases and the amount of noise added was decreased. At first, the noise was on the interval (-0.2,0.2), then the interval was reduced by 90% each time. Step 5 was applied as many times as needed to obtain a new total number of NNs equal to 50 (i.e., 50% of the remaining 100 NNs from the first stage). Finally, we construct the NN-based predictor as the ensemble of the remaining NNs from this second stage.

VI. Training of the Model

We applied the aforementioned techniques on the set of exemplars presented in section II.

First, we selected at random ten different sequences of compounds which formed ten *validation sets* of exemplars. These we named Sequence $1-10$.

For each sequence, the remaining 57 compounds, were supplied to the guided algorithm described in section IV which determined a *Training Set* and a *Test Set* of exemplars.

The heuristic described in section V was then applied to these sets to determine the response on the *Blind Set*. The results are in Table I. Sub-column "pIC_{50}" indicates the actual value of the Inhibitory Activity. The sub-column "Sensitivity Heuristic" refers to the difference between the actual and the predicted value provided by the model.

The compounds in the different *validation sets* are identified by a number; numbers $1-19$ represent compounds from the first family (nitrophenyl derivatives); numbers $20-46$ represent compounds from the second family (phenolic derivatives); and numbers $47-61$ represent compounds from the third family (pyridazine derivatives). Sub-column "N°" indicates the number of that compound and sub-column "Seq" indicates the number of a particular sequence.

The *Sensitivity Heuristic* produces responses that are very close to the expected ones, with a mean error (norm2) of 0.127 and a geometric mean of the relative errors of 0.052.

VII. Model Refinement

As outlined previously, given a set of compounds with known activities, our methodology partitions these compounds in different ways creating what we call *Sequences* of Neural Networks each of which models the activity of the compounds in the original set. These neural networks were so constructed so as to exhibit generalization abilities.

We shall use the term sequence to denote both the groups of the compounds that were used to develop Neural Networks that model the activity of the compound in the sequence as well as the set of Neural Networks that model this activity. In the following discussion, it will be clear in which context we shall use the term sequence.

For this experiment, we have created 10 partitions and their corresponding *Sequences* of trained Neural Networks. Each *Sequence* comprises 50 Neural Networks.

Collectively, we consider that those 500 Neural Networks, encompass the "knowledge" of the activity relation of the initial set of compounds (Aldose Reductase Inhibitory Activity).

TABLE I
MODEL RESPONSE TO VALIDATION SETS

Model	N°	pIC_{50}	Sensitivity Heuristic
1	*61*	9.07	2.72
	3	4.85	0.17
	43	5.93	0.82
	14	6.00	0.87
4	*32*	4.38	-0.78
	40	5.88	0.53
	22	5.18	-0.05
	34	5.50	0.26
7	*43*	5.93	0.51
	20	4.48	-1.18
	40	5.88	0.37
	6	4.74	-0.08
10	*47*	6.22	0.31
	23	4.86	-0.33
	12	5.62	0.24
	18	5.61	0.42

Model	N°	pIC_{50}	Sensitivity Heuristic
2	*57*	5.30	-1.30
	37	4.84	-0.06
	25	3.86	-1.36
	59	7.60	1.07
5	*26*	5.40	0.90
	18	5.61	0.01
	2	4.67	-0.34
	13	5.45	-0.01
8	*55*	6.85	0.40
	37	4.84	0.01
	58	6.82	0.29
	13	5.45	0.36

Model	N°	pIC_{50}	Sensitivity Heuristic
3	*10*	5.45	0.44
	53	6.72	-0.65
	24	4.83	0.46
	45	4.72	-0.94
6	*35*	4.96	-0.02
	37	4.84	-0.03
	47	6.22	0.20
	45	4.72	-0.90
9	*60*	8.60	2.15
	7	4.65	-0.12
	51	5.72	-0.22
	43	5.93	0.73

The issue now is on how to reach to an informed "decision" as to the activity of an unknown compound based on the "knowledge" incorporated in the set of the trained Neural Networks discussed earlier.

A method that we used in previous studies [16] and which yielded good results was to simply average the responses of the the sequences excluding the sequences that responded with the maximum and the minimum values respectively.

Recognizing that the response of some of the sequences may not be "accurate", we will not consider such sequences when we compute the response of the model of the unknown compound in question.

The question therefore is how to identify the sequences which provide a "non-accurate" response for the compound in question, given that the actual response of the compound is not known. The method outlined below, first establishes an *Initial Estimate* of the response. Expecting that the differences in responses of all Neural Networks are due to random rather than systematic effects, our method eliminates *Sequences* so that the spectrum of responses of the remaining Neural Networks is more evenly distributed above and below the *Initial Estimate*. The *Sequences* that remain are then used to calculate the response of the model.

We outline our procedure next.

Model Refinement Heuristic:

A. Establish the Initial Estimate

1) An array consisting of the value of the response of all the NNs to the compound in question is formed and sorted based on the value of the responses. Denote by $val_i; i = 1, \ldots N$ the i^{th} value in the sorted array.
2) The *Initial Estimate (IE)* is calculated as the mean of the 25th to 35th of the highest values and 25th to 35th of the lowest values.

$$IE = \left(\sum_{m=25}^{34} val_m + \sum_{n=N-34}^{N-25} val_n \right) /20 \qquad (4)$$

B. Eliminate Sequences

1) In Round0 (Preprocessing stage). The average value of the responses of all the Neural Networks in each *Sequence* is calculated.
2) In Round1, one calculates the difference of the values of the responses to the compound in question for all the Neural Networks in all the *Sequences* from the $InitialEstimate$.

$$ErrorPower = \sum (val_i - IE) \qquad (5)$$

One would expect that the $ErrorPower$ would be zero if the spectrum of responses were to be normal. If the $ErrorPower$ thus obtained is not zero, we eliminate the Sequence that has the smallest (largest) mean response if the $ErrorPower$ is positive (negative).
3) The previous step is applied five more times (Rounds 2, 3 , 4 , 5 and 6).

VIII. VALIDATION THROUGH BLIND TESTS

In order to validate the results of our approach, we conducted a blind test as follows. The collaborators at the Department of Pharmacy at the University of Thessaloniki, supplied the descriptors of nineteen compounds. These compounds, were anonymized (i.e. the names, structure and their biological activity (i.e. their pIC_{50}) were not disclosed.

TABLE II

THE COMPOUNDS USED IN THE BLIND VALIDATION SET

Type	Symbol	Actual Name
Active	AX-1	N-(3,5-difluoro-4-hydroxyphenyl)benzenesulfonamide
	AX-2	N-(3,5-difluoro-4-hydroxyphenyl)-4-methoxybenzenesulfonamide
	AX-3	N-(3,5-difluoro-4-hydroxyphenyl)-4-nitrobenzenesulfonamide
	AX-4	4-amino-N-(3,5-difluoro-4-hydroxyphenyl)benzenesulfonamide
	AX-5	N-(3,5-difluoro-4-hydroxyphenyl)-4-(1H-pyrrol-1-yl)benzenesulfonamide
	AX-6	N-(4-(N-(3,5-difluoro-4-hydroxyphenyl)sulfamoyl)phenyl)benzamide
	AX-7	N-(3,5-difluoro-4-hydroxyphenyl)-4-(trifluoromethyl)benzenesulfonamide
	AX-8	N-(3,5-difluoro-4-hydroxyphenyl)-4-(3-ethylureido)benzenesulfonamide
	AX-9	N-(4-(N-(3,5-difluoro-4-hydroxyphenyl)sulfamoyl)phenyl)-4-methoxybenzamide
	AX-18	N-(4-bromo-2-fluorobenzyl)-N-(3,5-difluoro-4-hydroxyphenyl)benzenesulfonamide
	AX-19	4-nitro-N-(4-bromo-2-fluorobenzyl)-N-(3,5-difluoro-4-hydroxyphenyl)benzenesulfonamide
	AX-20	4-amino-N-(4-bromo-2-fluorobenzyl)-N-(3,5-difluoro-4-hydroxyphenyl)benzenesulfonamide
	AX-21	N-(3,5-difluoro-4-hydroxyphenyl)-N-(phenylsulfonyl)benzenesulfonamide
Inactive	AX-10	N-((1H-tetrazol-5-yl)methyl)benzenesulfonamide
	AX-11	N-((1H-tetrazol-5-yl)methyl)-N-(phenylsulfonyl)benzenesulfonamide
	AX-13	N-(methylsulfonyl)-2-(phenylsulfonamido)acetamide
	AX-16	N-(1H-tetrazol-5-yl)benzenesulfonamide
Enantiomer	AX-15	(R)-N-(3-oxoisoxazolidin-4-yl)benzenesulfonamide
	AX-17	(S)-N-(3-oxoisoxazolidin-4-yl)benzenesulfonamide

TABLE III

COMPARISON BETWEEN THE PREDICTIONS RESULTED BY THE TWO METHODS AND THE ACTUAL RESULTS

Type	Compound	Experimental	Previous	Proposed Method
Active	AX-1	4.483	4.557	4.603
	AX-2	4.810	4.906	5.062
	AX-3	4.353	4.883	5.081
	AX-4	4.851	5.092	5.115
	AX-5	4.243	4.924	4.868
	AX-6	5.102	5.136	5.053
	AX-7	4.039	4.660	4.522
	AX-8	4.423	6.285	5.136
	AX-9	4.921	6.158	4.677
	AX-18	5.244	6.709	5.191
	AX-19	5.046	6.517	5.455
	AX-20	6.409	6.53	5.487
	AX-21	4.551	6.451	5.480
Inactive	AX-10	< 4	6.141	5.591
	AX-11	< 4	6.558	5.419
	AX-13	< 4	5.92	4.992
	AX-16	< 4	6.181	5.654
Enantiomer	AX-15	4.148	4.514	5.047
	AX-17	< 4	4.514	5.047

The previously obtained 10 ensembles of NN corresponding to the 10 sequences (Section VI) were used to derive the AR inhibitory activity of these unknown compounds.

The Model Refinement Heuristic presented in Section VII was used to obtain the predicted response.

We also obtained the predicted response as per the method used in our previous work [16].

The results appear in TABLE III and TABLE IV and Figures 1 and 2 while the *Sequences* used in predicting the activity of each component are shown in Figure 3.

TABLE IV
THE AVERAGE NORM VALUES FOR EACH METHOD

Type	Metric	Previous	Proposed Method
Active	Norm-2	0.289	0.148
	Norm-Inf	1.9	0.929
	Geo-Mean	0.089	0.065
Inactive	Norm-2	1.106	0.717
	Norm-Inf	2.558	1.654
	Geo-Mean	0.547	0.346
Enantiomer	Norm-2	0.316	0.691
	Norm-Inf	0.514	1.047
	Geo-Mean	0.106	0.238

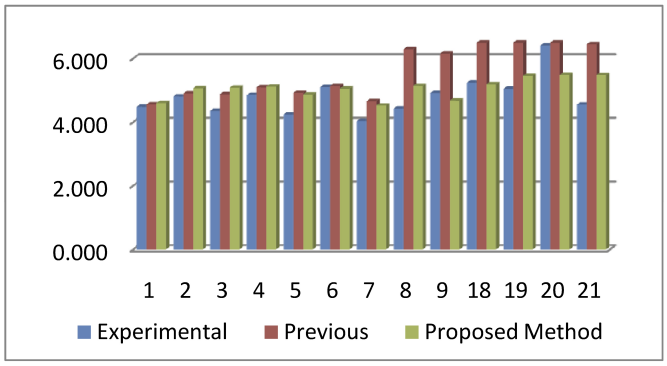

Fig. 1. Comparison Between the two Methods for Active Compounds

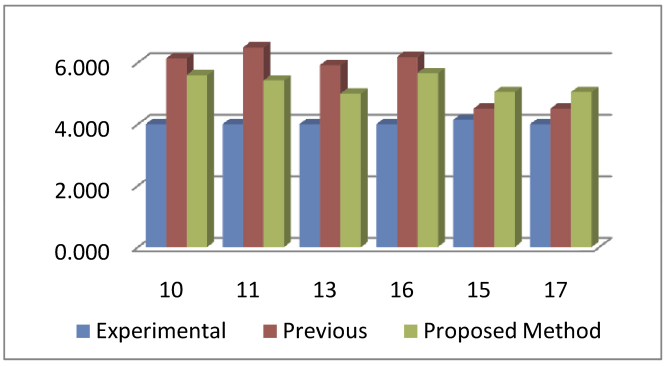

Fig. 2. Comparison Between the two Methods for Inactive and Enantiomer Compounds

The tables comprise two sections namely the section that includes compounds AX1 through 13 and the section with compounds AX 15 through 23. The compounds in the first section are fairly active compounds while the compounds in the second section are inactive (some of the activities were not determined experimentally and are reported as less than 4.0).

As it can be seen, our model behaves quite well, especially for the active compounds. It is markedly better than the model we reported in our previous work [16] and denoted as previous in the presented tables and figures. From TABLE IV we can see that on the average the error (i.e. the difference of

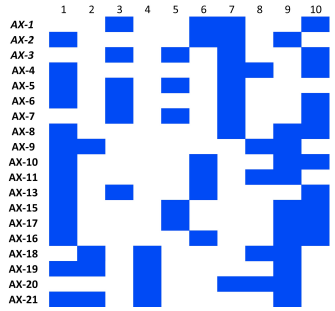

Fig. 3. The *Sequences* contributing to the prediction of the ARI activity of each compound

the response of proposed method from the experimentally obtained value) is 0.148 and the geometric mean of the relative errors is 0.065 while the maximum error is only 0.929. It is remarkable that these errors are very close to the ones we obtained for the set of the original compounds as per section VI.

The proposed method also improves significantly as compared to the method we presented previously. The error of the proposed method is nearly half the error obtained with the previous method. This is due to the selection of the sequences of Neural Networks that ensures that only relevant ones contribute to the final result.

These results are especially true for the set of the active compounds. For the set of inactive compounds, our method behaves well, but not as well as for the set of active compounds. The corresponding average and maximum errors are 0.717 and 1.654 respectively.

The reason for the observed discrepancy can be explained due to the fact that our model was only trained on active compounds and not on inactive ones. Given that the class of active compounds is much larger than that of the inactive ones, our model does not have enough examples of what constitutes an inactive compound to draw any sound inferences and hence generalize correctly.

Please note that it is difficult to accurately measure the activity of an inactive compound. The activity of such compounds

is often reported as less than 4.

It is significant that our model was trained on chemicals drawn from three distinct classes: namely nitrophenyl derivatives, phenolic derivatives, and pyridazine derivatives. The unknown compounds we used in this blind test did not belong to any of the previous three classes. Rather, they are sulfonamides, derivatives of the difluorophenol moiety. Yet, our model, was able to accurately predict the experimentally measured Aldose Reductase Inhibitory Activity of these unknown compounds. This leads us to believe that our method, is capable of deriving accurate models based on a very sparse set of exemplars, and that these models can generalize,and are able to accurately determine the activity of chemical compounds not related to the chemical compounds they were trained on.

IX. Conclusions and future work

In this work, we presented and validated a novel method of applying NN techniques in the area of QSAR. Our techniques are applicable to cases where the set of exemplars is sparse.

For both the training and validation trials, we used data that were laboratory measured and validated.

The results reported are significant not only because the errors are quite small, but more importantly because they are within a generally accepted absolute error of 1 [7].

Further we developed our model based on exemplars that represented three different families of chemical compounds, and validated it with chemical compounds from yet a fourth class of chemicals. Thus, our technique was able to produce a NN model that is accurate, and to the extend of the compounds it was tested on, can be considered *chemical-family-independent*.

Although the results we presented here very strongly indicate a NN model that can accurately predict the Aldose Reductase inhibitory activity of arbitrary chemical compounds, we feel that further validation trials, involving larger number of compounds, are necessary to fully explore the predictive abilities and limitations of the developed model.

Further, we plan to also utilize inactive compounds in our training set to ensure that our model can accurately generalize and predict the activity of inactive compounds.

We further plan to study the significance of each of the descriptors used (based on the *Sensitivity Heuristic*) and use this information together with a Principal Component Analysis to perhaps curtail the number of descriptors used.

Acknowledgment

Computational support was provided by the High Performance Computing Facility at the University of Victoria. This work was supported in part by The National Science and Engineering Research Council of Canada (NSERC) and by the Lansdowne Chair in Computer Engineering at the University of Victoria.

References

[1] Alexiou, P.; Nicolaou, I.; Stefek, M.; Kristl, A.; Demopoulos, V.J. Design and synthesis of N-(3,5-difluoro-4-hydroxyphenyl)benzenesulfonamides as aldose reductase inhibitors. Bioorg. Med.Chem., 2008, 16(7), 3926-32.

[2] Alexiou, P.; Pegklidou, K.; Chatzopoulou, M.; Nicolaou, I.; Demopoulos, V.J. Aldose Reductase Enzyme and its Implication to Major Health Problems of the 21st Century. Curr. Med. Chem., 2009, 16(6), 734-52.

[3] M. Ashton et al., "Identification of Diverse Database Subsets using Property-Based and Fragment-Based Molecular Descriptions," *Quant. Struct.-Act. Relat.*, Vol. 21, pp. 598−604, 2002.

[4] L. Costantino, A. M. Ferrari, M. C. Gamberini, G. Rastelli, "Nitrophenyl derivatives as aldose reductase inhibitors". Bioorg.Med.Chem. (2002), 10, 3923−3931.

[5] L. Costantino, A. Del Corso, G. Rastelli, J. M. Petrash, U. Mura, "7-Hydroxy-2-substituted-4-H-1-benzopyran-4-one as aldose reductase inhibitors: a SAR study". Eur.J.Med.Chem. (2001), 36, 697−703

[6] P. M. Granitto, P. F. Verdes, H. D. Navone, H. A. Ceccatto, "Aggregation algorithms for neural network ensemble construction", SBRN 2002. Proceedings VII Brazilian Symposium on Neural Networks, Nov. 2002, pp. 178−183

[7] P. Gillespie, R. A. Goodnow Jr., "The Hit-to-lead Process in Drug Discovery", in Annette M. Doherty (ed) *Annual Reports in Medicinal Chemistry* Vol. 39, pp. 293−321, 2004

[8] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selectio", Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1137−1143, San Francisco, CA, 1995. Morgan Kaufmann.

[9] L.K. Hansen, and P. Salamon, "Neural Network Ensembles", *IEEE Trans. Patrn. Anal. and Mchin. Intellignc.*, Vol. 12, No.10, 1990.

[10] E. M. Laxdal, R. Parra-Hernandez, and N. J. Dimopoulos, "Guided Construction of Training Data Set for Neural Networks". IEEE International Conference on Systems, Man & Cybernetics, October 2004, The Hague, The Netherlands, pp. 5905−5910.

[11] S. Miyamoto, "Recent Advances in Aldose Reductase Inhibitors: Potential Agents fot the Tratment of Diabetic Complications", *Expert Opin. Ther. Patents*, 12 , pp. 621−631, 2002.

[12] B. L. Mylari et al. "A highly selective, non-hydantoin, non-carboxylic acid inhibitor of aldose reductase with potent oral activity in diabetic rat models: 6-(5-chloro-3-methylbenzofuran-2-sulfonyl)-2-H-pyridazin-3-one". J.Med.Chem. (2003), 46, 2283−2286

[13] I. Nicolaou et al. "[1-(3,5-Difluoro-4-hydroxyphenyl)-1H-pyrrol-3-yl]phenylmethanone as a Biostere of a Carbrboxylic Acid Aldose Reductase Inhibitor", *Jrnl. of Medicinal Chmemistry*, 47, 2706-2709 (2004)

[14] Niculescu, Stefan P., "Artificial neural networks and genetic algorithms in QSAR," *Journal of Molecular Structure (Theochem)*, Vol. 622, pp. 71−83, 2003.

[15] Parra-Hernandez, R., E. M. Laxdal, N. J. Dimopoulos and P. Alexiou " A New Neural Network Ensemble Heuristic for a Predictor of the Aldose Reductase Inhibitory Activity", Proceedings ISPAS 2005 IEEE International Symposium on Signal Processing and Information, pp. 838−843, Athens, Greece, Dec. 2005.

[16] Parra-Hernandez, R., E. M. Laxdal, N. J. Dimopoulos, P. Alexiou and V. J. Demopoulos "Validation Results for a Neural Network Ensemble Predictor of Aldose Reductase Inhibitory Activity 18th EuroQSAR Symposium, Rhodes, September 2010.

[17] Tham, S. Y. and Agatonovic-Kustrin, S., "Application of the artificial neural network in quantitative structure-gradient elution retention relationship of phenylthiocarbamyl amino acids derivatives," *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 28, pp. 581−590, 2002.