Available online at www.sciencedirect.com
ScienceDirect
Carbohydrate Research 343 (2008) 1359–1365
Carbohydrate RESEARCH

Artificial neural network aided estimation of the electrochemical signals of monosaccharides on gold electrode

Fereydoon Gobal* and Amin Sadeghpour Dilmaghani

Department of Chemistry, Sharif University of Technology, PO Box 11365-9516, Tehran, Iran

Received 14 January 2008; received in revised form 2 March 2008; accepted 5 March 2008
Available online 13 March 2008

Abstract—Artificial neural networks were used to predict the oxidation peaks potentials of 7 monosaccharides under linear sweep voltammetry regime. Two sets of descriptors, one based on molecular properties calculated through DFT and another based on simple geometric distributions of hydroxyl groups and asymmetric carbon atoms along molecular chains, were employed to introduce the molecules to networks. Relatively, simple networks of (3,3,1) and (3,3,3,1) structures with the number of epochs not exceeding 15 through training process were capable of correctly predicting the peaks positions with R values in the range of 0.97–0.99.

Keywords: Artificial neural network; Monosaccharides; Electrochemical signals; Linear sweep voltammetry

1. Introduction

Monosaccharides are the building blocks of biologically and industrially important compounds and the investigation of their redox properties is of importance. Monosaccharides contain aldehyde, primary and secondary hydroxyl groups and the electrochemical reactivity of these functional groups strongly influence the total electro-oxidation behavior of the molecules. Electro-oxidation of monosacchaides[1] and some relevant compounds[2] on gold electrode in alkaline solution have been studied in detail aiming at the establishment of structure–reactivity relationship. In these studies, the method of linear sweep voltammetry has been employed and peaks positions in the current–potential plots have been correlated with the structures. Rather limited success has been met. The purpose of the present work is to use the method of artificial neural network to correlate the peak position in anodic linear sweep voltammograms of monosaccharides electro-oxidation to their fundamental properties.

Artificial neural network (ANN) arises from attempts to mimic the functioning of human brain and is in principal capable of firstly learning and secondly predicting the relations between sets of factors and responses even though there is no theoretically defined inter-relations or relations are nonlinear.[3] An ANN consists of a number of neurons (nodes) distributed into input, hidden, and output layers and there are connections between layers with weights signifying the strength of the signal flowing into the nodes. The response of a node, the signal it sends down the layers, is controlled by comparing its output against some pre-set function that normally is a sigmoid. Having provided the input, the algorithm sets the weights so that the output is within an acceptable errors limit of the experimental output (supervised learning). In this work, feed forward back propagation algorithm with a tansigmoid activation transfer function has been used.[3]

2. Methods

To introduce the monosaccharides to a neural network, proper descriptors sensibly signifying their characteristics are required. The mission of ANN would then be to express the oxidation peak potential obtained under the regime of linear sweep voltammetry (LSV) in terms of some function of the descriptors.

* Corresponding author. Tel.: +98 21 66005718; fax: +98 21 66012983; e-mail: Gobal@sharif.ir

0008-6215/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.carres.2008.03.012

The LS voltammograms of monosaccharides, Figure 1a and b, were obtained from the literature[1] where only the first potential sweep data have been considered due to the changes experienced upon repeated sweeps that are probably due to the fouling of the electrode. Potential sweeps in the range of 0–1.5 V/RHE and potential sweep rates of 5–50 mV s^{-1} have been employed. According to the existing literature three regions in the votammograms of monosaccharides containing all electrochemically active functional groups can be resolved at 50 mV s^{-1} sweep rates. In region A ($E < 0.9$ V/RHE), the electro-oxidation of aldehyde hydrate (or hemiacetal) at carbon no. 1 occurs. In region B ($0.9 < E < 1.1$ V/RHE) the current peak is due to the oxidation of primary hydroxyl groups. Vicinal diols oxidation leading to the cleavage of C–C bond occurs in the range of 1.1–1.3 V/RHE and named region C. Understandably, the peaks tend to merge as the potential sweep rates are increased. Figure 1a and b present the voltammograms of 7 monosaccharides.[1]

Asymmetric peaks in the range of 0.9–1.5 V/RHE have been resolved to two peaks while three overlapping peaks of arabinose in the range of 0.7–0.9 V/RHE are merged to form a broad and asymmetric peak (similar to other monosaccharides) suitable for ANN analysis. Table 1 presents the ultimate peaks positions.

As for the selection of descriptors the method of density function theory (DFT) has been employed to first optimize the structures, then to calculate the energetic and electronic properties of the sugar molecules. In this regard, DFT calculation based on B3LYP algorithm was performed using 6-31G(d) basis set with GAUSSIAN 98 software.[4] A computer program based on Levenberg–Marquardt algorithm was written in Matlab

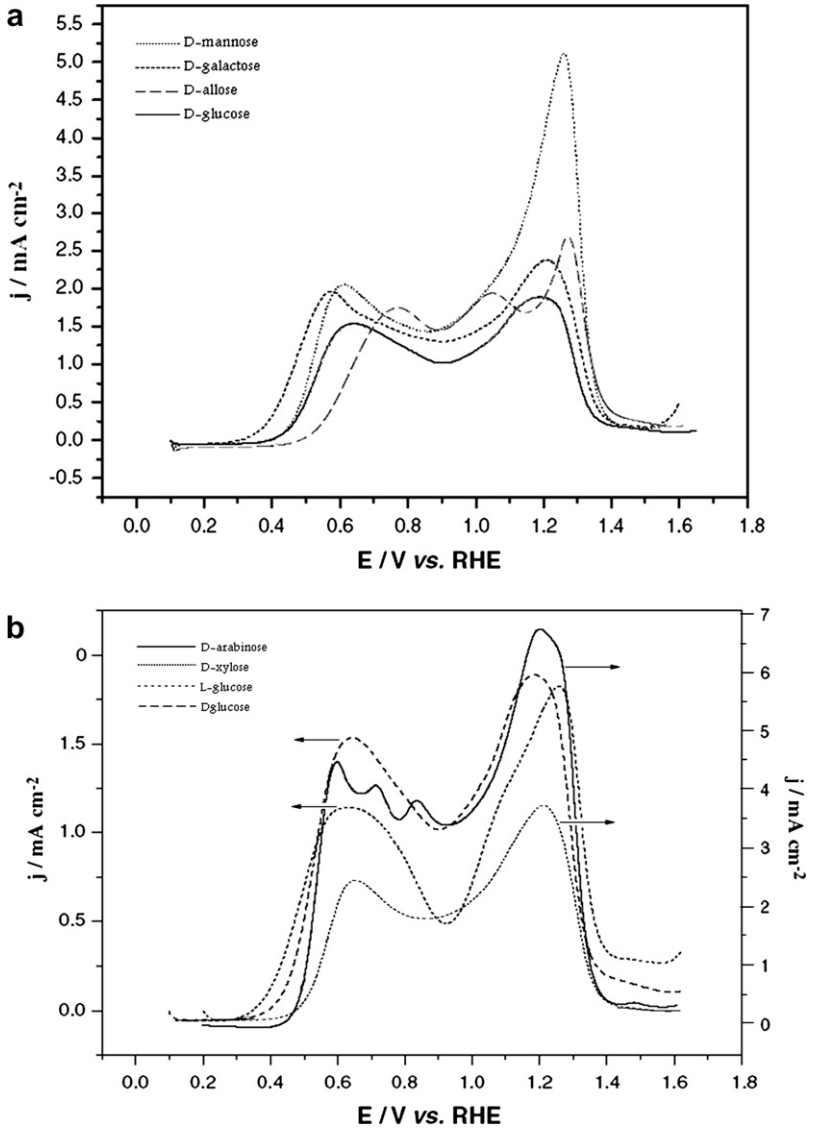


Figure 1. (a) Anodic linear sweep voltammograms of gold in the presence of 10 mM D-allose, D-glucose, D-galactose, D-mannose in 0.1 M NaOH. $v = 5$ mV s^{-1}, 25 °C; (b) in the presence of 10 mM D-glucose, L-glucose, D-arabinose, D-xulose in 0.1 M NaOH. $v = 5$ mV s^{-1}, 25 °C (Ref. 1).

Table 1. The electrochemical signals of different monosaccharides obtained from the literature[1]

	First peak	Second peak	Third peak
1-D-Allose	0.768	1.045	1.298
2-D-Glucose	0.635	1.100	1.270
3-D-Galactose	0.575	1.230	1.280
4-D-Arabinose	0.680	1.103	1.255
5-D-Mannose	0.608	1.177	1.284
6-L-Glucose	0.643	1.145	1.295
7-D-Xylose	0.662	1.150	1.300

environment in this laboratory to perform ANN calculations.

3. Results and discussion

As a typical output the optimized structure of D-glucose and the components of its dipole moment is presented in Figure 2. Obviously the orientation of dipole moment depends on how hydroxyl groups are placed in either side of the molecule which in turn affects its interaction with the electrode surface and subsequent oxidation.

Dipole moments are selected as the first descriptors. As for the selection of energetic descriptors we resort to the model of Gerischer and Mindt[5] who considered the overlap of the electronic states of solutions entities and the electronic bands of the electrode in the course of electron transfer processes. Assuming that the concentrations of Ox and Red forms are equal, the Fermi level in solution lies half way between the electronic levels of Red and Ox species in solution and is assumed to be fixed while the Fermi level of the metallic electrode can be monitored (raised or lowered) by the imposed potential. Raising the Fermi level enhances reduction and its lowering promotes oxidation of the solution species. Under these criteria the positions of the highest occupied molecular orbitals (HOMO) of different monosaccharides were calculated by the B3LYP/6-31G(d) algorithm. The HOMO levels computed in this way are a good approximation of the first ionization potential as a consequence of the interplay between the Koopmans recipe and the Hohenberg–Kohn (sham) theorem.[6] This also in a sense reflects the electro-oxidation potential and is taken as the second descriptor. The method also provides the sum of electronic and thermal free energies, total free energy of molecule, which surely depends on the number of constituent atoms. We have normalized the values by dividing by the number of carbon atoms and used as the third descriptor. Table 2 presents the values of descriptors used to introduce the monosaccharides to the network.

An artificial neural network based on the back propagation of error with tansigmoid activation transfer functions in the hidden layer(s) and a linear transfer function in the output layer was employed. It was found that a three layer network having 3 nodes in the input, 2

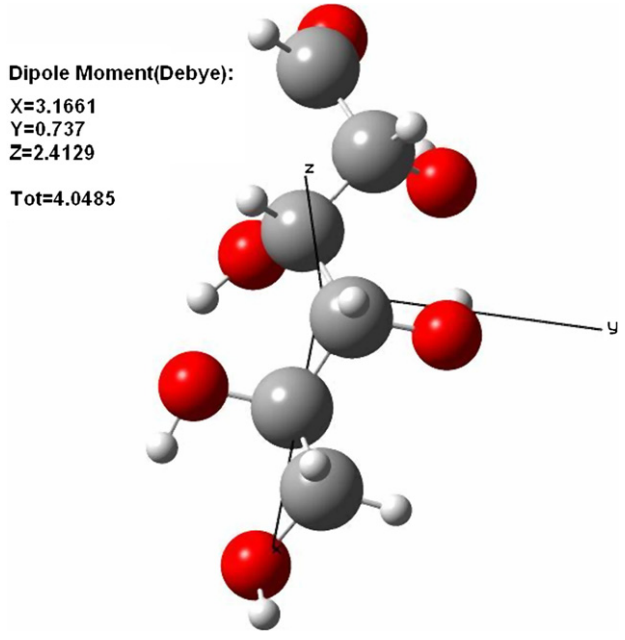


Figure 2. Optimized structure of D-glucose calculated by DFT(B3LYP) using G98-A.

Table 2. Calculated properties of monosaccharides that are used as electrostatic descriptors

	HOMO (a.u)	Modified total energy (a.u)	Dipole moment (D)
1-D-Allose	−0.258	−114.49803	3.079
2-D-Glucose	−0.257	−114.49689	4.049
3-D-Galactose	−0.251	−114.49647	4.074
4-D-Arabinose	−0.261	−114.49631	5.200
5-D-Mannose	−0.250	−114.49723	3.286
6-L-Glucose	−0.258	−114.49754	3.659
7-D-Xylose	−0.262	−114.49716	3.990

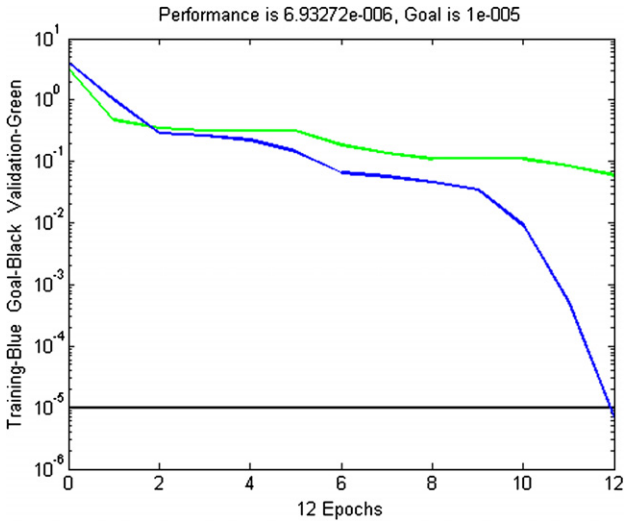


Figure 3. Presentations of errors in the course of training and validation processes via Levenberg–Marquardt algorithm using electrostatic descriptors.

nodes in the hidden and one node in the output layer and having learning rates and momentum of 0.5 and 0.9 in the optimized network is well capable of predicting peaks positions in the LSV pattern. The number of iterations before the error starts to rise was about 12, (Fig. 3). Due to the smallness of the data set, cross validation through the use of Leave-One-Out[7] has been employed to take care of the small size of the data set.

Figure 4a–c present the correlations of the experimental and ANN predicted peak positions. *R* values in the range of 0.974–0.985 signify goodness of the fits and the prediction capability of the network. Also, and more importantly is the fact that the oxidation peak potentials can be somehow expressed in terms of more fundamental and molecular properties of the compounds.

The above descriptors however do not explicitly take into account the structure of the sugar molecules so far as the distribution of hydroxyl groups in the molecule as well as the chirality of carbon atoms are concerned. Thus, the more distinct and precise characteristics of molecules must be selected as descriptors. In this regard, the distance between the oxygen atoms bound to carbons nos. 2 and 3 (D23) affects the extent of tautomerization (aldehyde hemiacetal equilibrium) occurring on carbon no. 1 and is selected as a descriptor. Also, the distance between two nearest hydroxyl groups in the

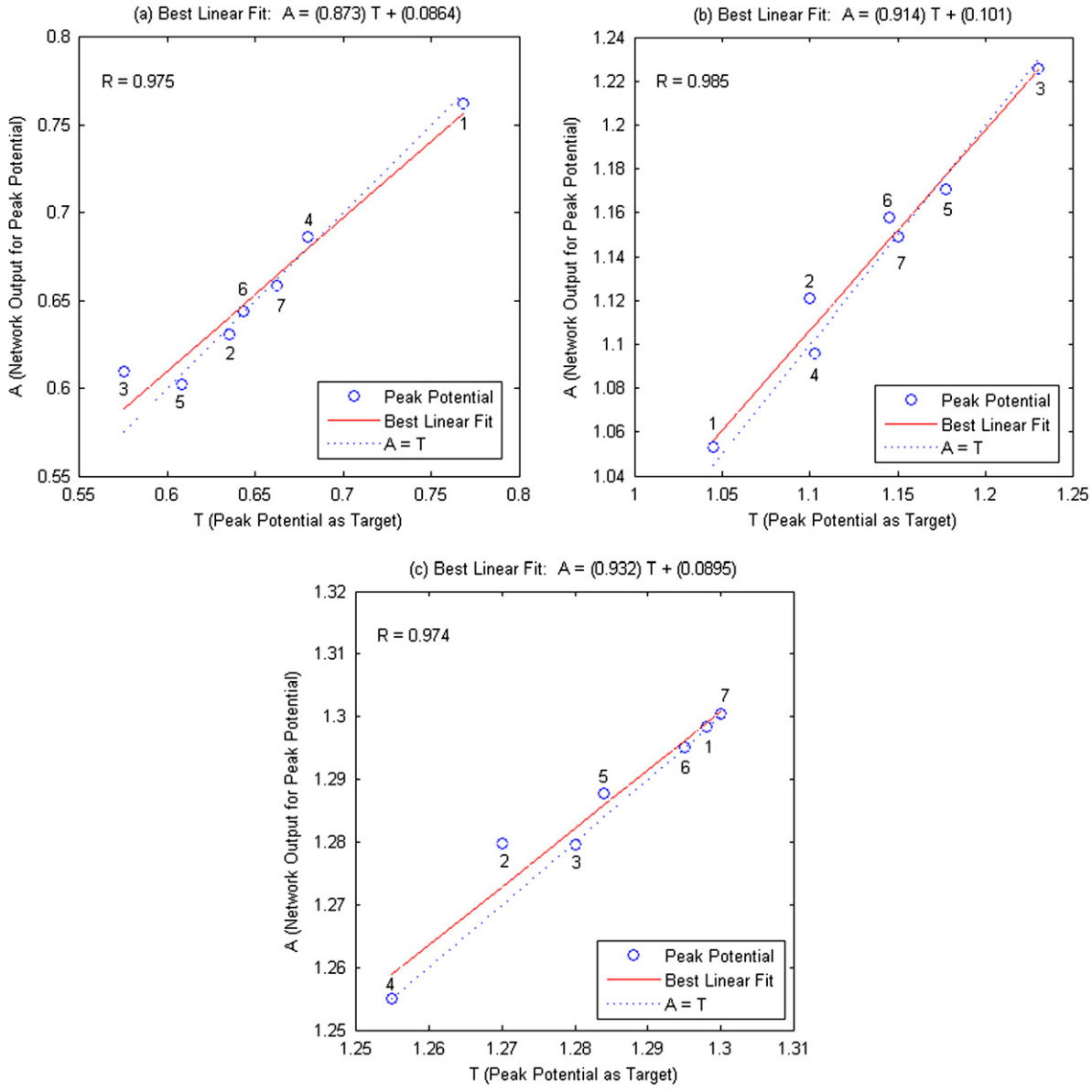


Figure 4. Calculated versus observed peak potential values of (a) aldehyde oxidation with learning rate and momentum constant values of 0.50 and 0.90; (b) primary alcohol oxidation with learning rate and momentum constant values of 0.58 and 0.50; (c) vicinal diols oxidation with learning rate and momentum constant values of 0.57 and 0.90 for LOO sets from the 3 electrostatic descriptors model after submission to computational neural network.

entire structure (Dn) that affects the oxidation potential of the vicinal groups is taken as the second descriptor. Chirality is a characteristic of sugars and could be taken as a good descriptor which can indeed discriminate between stereoisomers.

Topological indices have been developed for stereoisomeric organic compounds which can identify the presence of stereocenters in molecules.[8] In this regard, Schultz et al.[9] has introduced a chiral factor (CF) that represents the *R* configuration as +1 and the *S* configuration of chiral carbon as −1 with CF value of zero is assigned to nonstereocenters. We have arbitrary defined the total chirality of molecule by multiplying two vectors. The first is a 1 × 6 vector for a six membered sugar that we named weighting vector where the elements $m_i = 1/i$ and i is the carbon number starting from the aldehyde side when estimating the aldehyde oxidation peak potential, and from the alcohol side when estimating the peak potential due to the oxidation of the primary alcohol group. This vector in a sense signifies the distances of the chiral (or else) centers to the electroactive center. The elements of second vector represent the chirality of the carbon atoms. Accordingly, for D-glucose the descriptors are calculated as

$$[0 \quad 1 \quad -1 \quad 1 \quad 1 \quad 0] \times \begin{bmatrix} 1.00 \\ 0.50 \\ 0.33 \\ 0.25 \\ 0.20 \\ 0.17 \end{bmatrix} = 0.62$$

for the aldehyde peak potential estimation (CFald) and

$$[0 \quad 1 \quad 1 \quad -1 \quad 1 \quad 0] \times \begin{bmatrix} 1.00 \\ 0.50 \\ 0.33 \\ 0.25 \\ 0.20 \\ 0.17 \end{bmatrix} = 0.79$$

for the primary alcohol peak potential estimation (CFalc). Table 3 presents the descriptors for the sugars used in this study.

The first two oxidation peak potentials could be predicted using D23, Dn, and one of the CFald or CFalc for the first and second oxidation peaks, respectively, through a neural network comprising of 3 nodes in the input layer, and two hidden layers each having 3 nodes (with tansigmoidal transfer functions) and one output layer (with linear transfer function). The correlation results are presented in Figure 5a and b where *R* values of 0.988 and 0.969 signify good correlations.

Learning rates of 0.9 and 0.5 and momentum values of 1.1 and 0.3 have been used in the optimized nets for the prediction of the oxidation peaks of the aldehyde and primary alcohol groups. Attempts to predict the oxidation peak potential of the vicinal hydroxyl groups

Table 3. Structural and topological descriptors

	Dn	D23	CFald	CFalc	CrF
1-D-Allose	3.61	3.63	1.28	1.28	−4.66
2-D-Glucose	2.83	2.83	0.62	0.78	1.32
3-D-Galactose	3.00	3.02	0.12	0.12	−0.66
4-D-Arabinose	2.86	2.86	−0.12	−0.58	0.00
5-D-Mannose	3.27	3.61	−0.38	0.38	3.32
6-L-Glucose	2.83	2.94	0.22	−0.22	8.66
7-D-Xylose	2.72	2.83	0.28	0.42	5.00

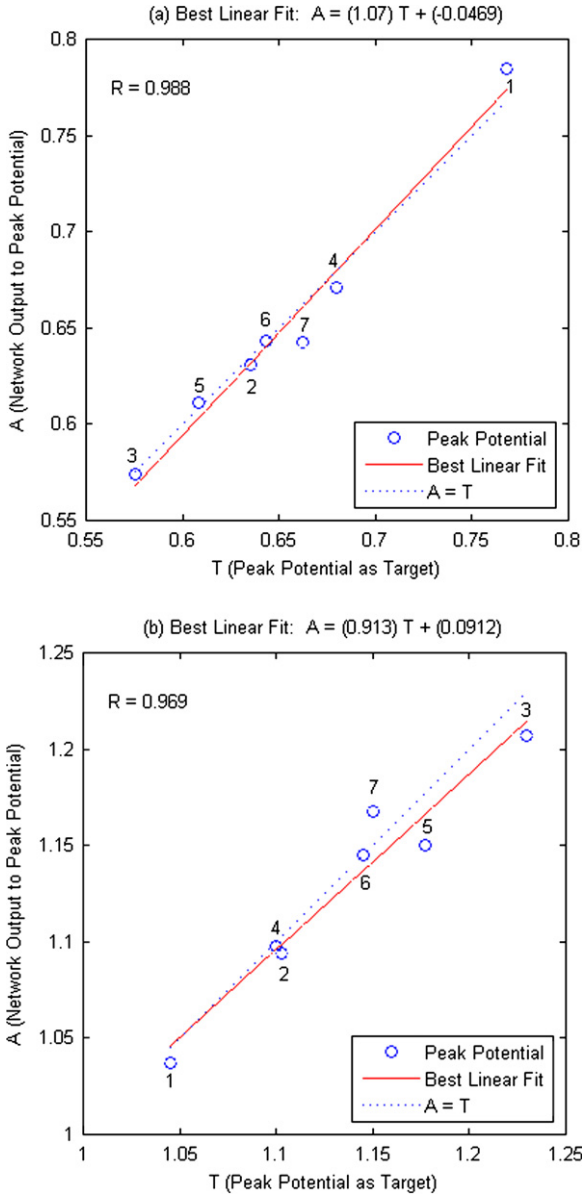
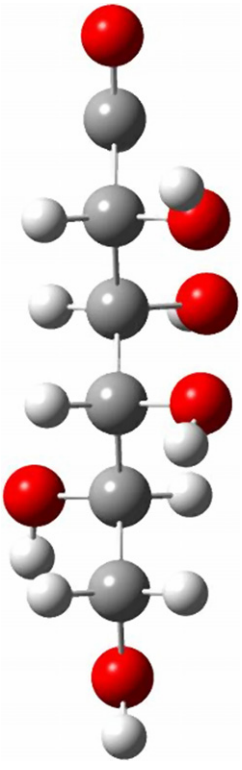


Figure 5. (a) Calculated versus observed peak potential values of aldehyde oxidation with learning rate and momentum constant values of 0.90 and 1.10; (b) primary alcohol oxidation with learning rate and momentum constant values of 0.50 and 0.30, for LOO sets from the 3 topological descriptors model after submission to computational neural network.

were not successful and prompted the introduction of yet another topological descriptor capable of considering relative positions of the hydroxyl groups. To create such a descriptor one has to consider both the orientation (same side or opposite side of the chain of carbon backbone) and the separation distance of the hydroxyl groups. To achieve this, a characteristic matrix with elements $c(i,j)$ with i and j signifying the number of carbon atoms bearing the hydroxyl groups is generated. If both hydroxyl groups are on the same carbon atom, $c(i,j)$ is set to zero while hydroxyl groups on the same side of the chain of carbon atoms are given $+1$ and on the opposite sides of the chain are given -1. For D-glucose with the structure presented in Scheme 1 and numbering the optically active carbon atoms from the aldehyde side the characteristic matrix is

$$c = \begin{bmatrix} 0 & +1 & +1 & -1 \\ +1 & 0 & +1 & -1 \\ +1 & +1 & 0 & -1 \\ -1 & -1 & -1 & 0 \end{bmatrix}$$

In this matrix representation $c(1,4)$ with the value of -1 signifies that the hydroxyl groups on carbons 1 and 4 are on either side of the molecular chain. Similarly, $c(2,3)$ with the value of $+1$ shows that the hydroxyl groups on the carbon atoms number 2 and 3 are both located at the same side of the molecular chain.



Scheme 1.

A second matrix with $w(i,j) = 1/|i-j|$ elements is created to account for the separations of the hydroxyl groups on carbon no i and j

$$w = \begin{bmatrix} 0 & 1 & 1/2 & 1/3 \\ 1 & 0 & 1 & 1/2 \\ 1/2 & 1 & 0 & 1 \\ 1/3 & 1/2 & 1 & 0 \end{bmatrix}$$

where $w(1,2)$ states that the separation is twice as long compared to $w(1,3)$.

Finally, multiplying the two creates a matrix that combines both the orientation and separation measures where for D-glucose is

$$[\mathrm{CrF}] = \begin{bmatrix} 0 & 1 & 1/2 & 1/3 \\ 1 & 0 & 1 & 1/2 \\ 1/2 & 1 & 0 & 1 \\ 1/3 & 1/2 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 0 & +1 & +1 & -1 \\ +1 & 0 & +1 & -1 \\ +1 & +1 & 0 & -1 \\ -1 & -1 & -1 & 0 \end{bmatrix} = \begin{bmatrix} +1.17 & +0.17 & +0.67 & -1.50 \\ +0.50 & +1.50 & -0.50 & -2.00 \\ 0.00 & -0.50 & +0.50 & -1.50 \\ +1.50 & +1.17 & +0.83 & -1.83 \end{bmatrix}$$

and we define 'cross factor', CrF, as the trace of the matrix.

$$\mathrm{CrF}(\text{D-glucose}) = \mathrm{Trace}[\mathrm{CrF}] = 1.33$$

This factor and the two explained before, Dn and D23, are used to predict the third peak potential that is due to the oxidation of vicinal diols. A network

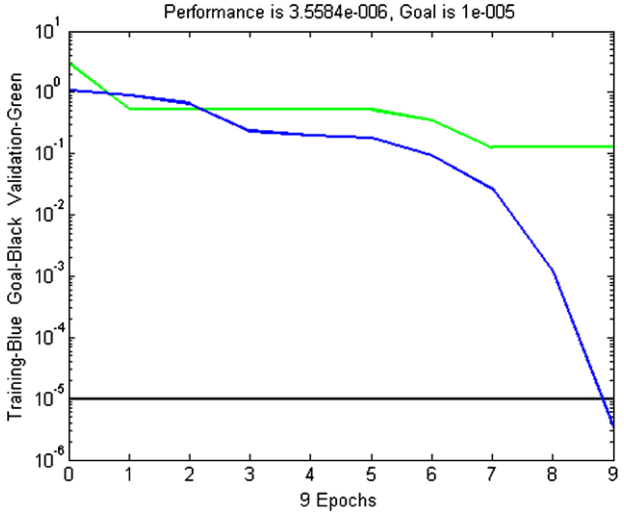


Figure 6. Presentations of errors in the course of training and validation processes via Levenberg–Marquardt algorithm using topological descriptors.

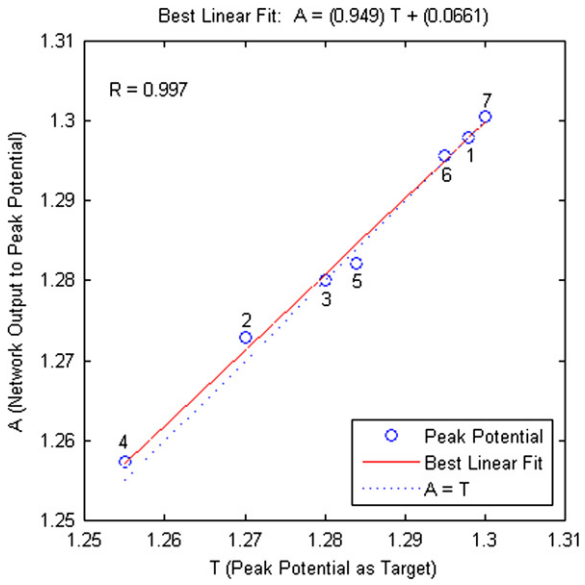


Figure 7. Calculated versus observed peak potential values of vicinal diols oxidation with learning rate and momentum constant values of 0.30 and 1.50.

comprising of 3 nodes in the input layer, and two hidden layers each having 3 nodes and one in the output layer with learning rate and momentum values of 0.3 and 1.5 has been employed. Tansigmoidal transfer function was used in the two hidden layers and linear transfer function was employed in the output layer. The number of epoch to achieve this result was around 9, Figure 6. Correlation results are presented in Figure 7.

4. Conclusion

Based on the above findings it is concluded that neural networks are well capable of predicting the electro-oxidation peaks potentials of monosaccharides in a linear potential sweep regime if sensible descriptors of molecules are defined. Geometric as well as energetic criteria may equally well be employed. The correlations are in the practice good to excellent.

References

1. Parpot, P.; Nunes, N.; Bettencourt, A. P. *J. Electroanal. Chem.* **2006**, *596*, 65–73.
2. Anastasijevic, N. A.; Baltruschat, H.; Heitbaum, J. *Electrochim. Acta* **1993**, *38*, 1067–1072.
3. Patterson, D. W. *Artificial Neural Networks: Theory and Applications*, 1st ed.; Prentice Hall: Singapore, 1996, pp 141–243.
4. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A. J.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A.; GAUSSIAN 98 (Revision A.6), Inc., Pittsburgh, PA, 1998.
5. Gerischer, H.; Mindt, W. *Surf. Sci.* **1966**, *4*, 431–439.
6. Bickelhaupt, F. M.; Baerends, E. J. *Rev. Comp. Chem.* **2000**, *15*, 1–86.
7. Bishop, C. M. *Neural Networks for Pattern Recognition*, 1st ed.; Clarendon Press: Oxford, 1995, pp 372–375.
8. Zhang, Q. Y.; Sousa, J. A. *J. Chem. Inf. Model* **2006**, *46*, 2278–2287.
9. Schultz, H. P.; Schultz, E. B.; Schultzs, T. P. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 864–870.