

## Accelerated Articles

# Use of Artificial Neural Networks for the Accurate Prediction of Peptide Liquid Chromatography Elution Times in Proteome Analyses

Konstantinos Petritis, Lars J. Kangas, Patrick L. Ferguson, Gordon A. Anderson, Ljiljana Paša-Tolić, Mary S. Lipton, Kenneth J. Auberry, Eric F. Strittmatter, Yufeng Shen, Rui Zhao, and Richard D. Smith\*

Biological Sciences Division and Environmental and Molecular Laboratory, Pacific Northwest National Laboratory, P.O. Box 999, Richland, Washington 99352

**The use of artificial neural networks (ANNs) is described for predicting the reversed-phase liquid chromatography retention times of peptides enzymatically digested from proteome-wide proteins. To enable the accurate comparison of the numerous LC/MS data sets, a genetic algorithm was developed to normalize the peptide retention data into a range (from 0 to 1), improving the peptide elution time reproducibility to ~1%. The network developed in this study was based on amino acid residue composition and consists of 20 input nodes, 2 hidden nodes, and 1 output node. A data set of ~7000 confidently identified peptides from the microorganism *Deinococcus radiodurans* was used for the training of the ANN. The ANN was then used to predict the elution times for another set of 5200 peptides tentatively identified by MS/MS from a different microorganism (*Shewanella oneidensis*). The model was found to predict the elution times of peptides with up to 54 amino acid residues (the longest peptide identified after tryptic digestion of *S. oneidensis*) with an average accuracy of ~3%. This predictive capability was then used to distinguish with high confidence isobar peptides otherwise indistinguishable by accurate mass measurements as well as to uncover peptide misidentifications. Thus, integration of ANN peptide elution time prediction in the proteomic research will increase both the number of protein identifications and their confidence.**

Proteomics involves the broad and systematic analysis of proteins, which includes their identification, quantification, and ultimately the attribution of one or more biological functions.<sup>1–3</sup> Proteomic analyses are challenging because of the high complexity and dynamic range of protein abundances. The industrialization of biology requires that the systematic analysis of expressed proteins be conducted in a high-throughput manner and with high sensitivity, further increasing the challenge. Recent technological advances in instrumentation, bioinformatics and automation have contributed to this goal. Specifically, in the area of proteomics, it is evident that greater specificity benefits the ability to deal with the high complexity of proteomes.<sup>4,5</sup> As a result, recent efforts have focused on improvements in separation speed, resolving power, and dynamic range, and these methods have generally been based on the combination of separations with mass spectrometry (MS), using correlation of tandem mass spectra with established protein databases or predictions from genome sequence data for identifications.<sup>6–8</sup>

At the present, there are two major approaches for proteomic analyses. The first one consists of the off-line combination of two-

- (1) James, P. *Proteome Research: Mass Spectrometry*; Springer: Berlin, 2001.
- (2) Wilkins, M. R.; Williams, K. L.; Appel, R. D. Hochstrasser, D. F. *Proteome Research: New Frontiers in Functional Genomics*; Springer: Berlin, 1997.
- (3) Yates, J. R., III. *J. Mass Spectrom.* **1998**, *33*, 1–19.
- (4) Peng, J.; Gyri, S. P. *J. Mass Spectrom.* **2001**, *36*, 1083–1091.
- (5) Unger, J.; Racaiyte, K.; Wanger, K.; Miliotis, T.; Edholm, L. E.; Bischoff, R.; Marko-Varga, G. *J. High Resolut. Chromatogr.* **2000**, *23*, 259–265.
- (6) Gygi, S. P.; Corthals, G. P.; Zhang, Y.; Rochon, Y.; Aebersold, R. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 9390–9395.
- (7) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., III. *Nat. Biotechnol.* **1999**, *17*, 676–682.
- (8) Smith R. D.; Anderson, G. A.; Lipton, M. S.; Pasa-Tolic, L.; Shen, Y.; Conrads, T. P.; Veenstra, T. D.; Udseth, H. R. *Proteomics* **2002**, *2*, 513–523.

\* Corresponding author.

dimensional polyacrylamide electrophoresis (2D-PAGE) with MS.<sup>6,9,10</sup> The proteins are first separated in a gel by their pI and mass, and then the protein "spots" are enzymatically hydrolyzed, resulting in peptide mixtures which can be analyzed by matrix-assisted laser desorption ionization-time-of-flight (MALDI-TOF) or electrospray (ESI)-MS.<sup>7,8,11–12</sup> Another rapidly evolving approach consists of a global proteome-wide enzymatic digestion followed by analysis using on-line 1-D or 2-D liquid chromatography (LC) coupled with ESI-MS. The detection and identification of the peptides is achieved by tandem MS<sup>7,13</sup> or more recently, by accurate mass measurements (e.g., using Fourier transform ion cyclotron resonance (FTICR)-MS).<sup>8,11,12,14,15</sup>

An aspect of proteomic analysis that has not yet been well-exploited involves use of the information available from LC elution or retention times. Indeed, retention time in LC is characteristic and structurally dependent for a defined experiment (mobile phase composition, stationary phase, etc.). If there were a way to predict the LC retention time for a given peptide structure, then this could be used in conjunction with either MS/MS data to improve the confidence of peptide identifications or to increase the number of peptide identifications, or with sufficiently high accuracy mass measurements, to reduce the need for MS/MS data (i.e., if the combination of elution time and mass accuracy provides sufficient specificity).

The idea that chromatographic behavior of peptides could be predicted on the basis of the amino acid composition is not new. In 1951, Knight<sup>16</sup> and Pardee<sup>17</sup> showed that synthetic peptide retention factor ( $R_f$ ) values on paper chromatography could be predicted with some accuracy. In 1952, Sanger<sup>18</sup> addressed the problem of isomers by demonstrating that the relationship between  $R_f$  and composition was not absolutely accurate, since peptides containing the same amino acids but having different sequences could frequently be separated. More recently, there have been several reports on the prediction of peptide elution times in reversed-phase (RP)<sup>19–27</sup> or normal phase<sup>28,29</sup> liquid

chromatography. These methods used quantitative structure–chromatographic retention relationships (QSRR's) (e.g., partial least-squares or multiple linear regression) for the peptide elution time prediction. Casal et al.<sup>30</sup> demonstrated that partial least-squares regression provides a better predictive ability with these models using a mixture of 25 small standard peptides. One limitation of these models is that they are most effective for peptides with fewer than 15–20 amino acid residues. Very recently, Palmblad et al.<sup>31</sup> used the predictive capability of the model previously described<sup>19,32,33</sup> (based on the summation of the coefficients of all amino acid residues) for the prediction of retention times for tryptic peptides. However, using tryptically digested commercial proteins as their training set, the accuracy of the prediction achieved was relatively poor as compared to that reported previously.<sup>31–33</sup>

Another approach based on artificial neural networks (ANNs) has demonstrated better predictive capabilities in several areas of chemistry, including (i) conformational states for small peptides,<sup>34</sup> (ii) carbon-13 nuclear magnetic resonance chemical shifts,<sup>35</sup> and (iii) the retention factor or retention time of small molecules in thin-layer chromatography,<sup>36</sup> GC,<sup>37–39</sup> and LC.<sup>40–42</sup> To our knowledge, other groups have not yet used ANNs for peptide elution time prediction. Several reviews of strategies for prediction of retention in LC as well as the application of ANNs in chemistry have been published elsewhere.<sup>43–45</sup>

In this work, we describe the use of an ANN for the prediction of peptide LC elution times. The development of our initial ANN model was based on the assumption that peptide elution times should substantially depend on amino acid compositions. The

- (9) Tonella, L.; Hoogland, C.; Blinz, P. A.; Appel, R. D.; Hochstrasser, D. F.; Sanchez, J. C. *Proteomics* **2001**, *1*, 409–423.
- (10) Fountoulakis, M.; Langenn, H.; Evers, S.; Gray, C.; Takacs, B. *Electrophoresis* **1997**, *18*, 1193–1202.
- (11) Shen, Y.; Tolić, N.; Zhao, R.; Paša-Tolić, L.; Li, L.; Berger, S. J.; Harkewicz, R.; Anderson, G. A.; Belov, M. E.; Smith, R. D. *Anal. Chem.* **2001**, *73*, 3011–3021.
- (12) Shen, Y.; Zhao, R.; Belov, M. E.; Conrads, T. P.; Anderson, G. A.; Tang, K.; Paša-Tolić, L.; Veenstra, T. D.; Lipton, M. S.; Udseth, H. R.; Smith, R. D. *Anal. Chem.* **2001**, *73*, 1766–1775.
- (13) Washburn, M. P.; Wolters, D.; Yates, J. R., III. *Nat. Biotechnol.* **2001**, *19*, 242–247.
- (14) Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Masselon, C.; Paša-Tolić, L.; Shen, Y.; Udseth, H. R. *OMICS* **2002**, *6*, 61–90.
- (15) Lipton, M. S.; Paša-Tolić, L.; Anderson, G. A.; Anderson, D. J.; Auberry, D. L.; Battista, J. R.; Daly, M. J.; Fredrickson, J.; Hixson, K. K.; Kostandarithes, H.; Masselon, C. D.; Markille, L. M.; Moore, R. J.; Romine, M. F.; Shen, Y.; Tolić, N.; Udseth, H. R.; Venkateswaran, A.; Wong, K. K.; Zhao, R.; Smith, R. D. *Proc. Natl. Acad. Sci.* **2002**, *99*, 11049–11054.
- (16) Knight, C. A. *J. Biol. Chem.* **1951**, *190*, 753–756.
- (17) Pardee, A. B. *J. Biol. Chem.* **1951**, *190*, 757–762.
- (18) Sanger, F. *Adv. Protein Chem.* **1952**, *7*, 1–7.
- (19) Meek, J. L. *Proc. Natl. Acad. Sci. U.S.A.* **1980**, *77*, 1632–1636.
- (20) Meek, J. L.; Rossetti, Z. L. *J. Chromatogr.* **1981**, *211*, 15–28.
- (21) Browne, C. A.; Bennett, H. P. J.; Solomon, S. *Anal. Biochem.* **1982**, *124*, 201–208.
- (22) Guo, D.; Mant, C. T.; Taneja, A. K.; Parker, J. M. R.; Hodges, R. S. *J. Chromatogr.* **1986**, *359*, 499–517.
- (23) Mant, C. T.; Burke, T. W. L.; Black, J. A.; Hodges, R. S. *J. Chromatogr.* **1988**, *458*, 193–205.

- (24) Wilce, M. C. J.; Aguilar, M. I.; Hearn, M. T. W. *J. Chromatogr.* **1991**, *536*, 165–183.
- (25) Wilce, M. C. J.; Aguilar, M. I.; Hearn, M. T. W. *J. Chromatogr.* **1993**, *632*, 11–18.
- (26) Sanz-Nebot, V.; Toro, I.; Barbosa, J. *J. Chromatogr., A* **2001**, *933*, 45–56.
- (27) Mills, M. J.; Maltas, J.; Lough, W. J. *J. Chromatogr., A* **1997**, *759*, 1–11.
- (28) Yoshida, T. *J. Chromatogr., A* **1998**, *811*, 61–67.
- (29) Yoshida, T.; Okada, T. *J. Chromatogr., A* **1999**, *841*, 19–32.
- (30) Casal, V.; Martin-Alvarez, P. Z.; Herraz, T. *Anal. Chim. Acta* **1996**, *326*, 77–84.
- (31) Palmblad, M.; Ramström, M.; Markides, K. E.; Håkansson, P.; Bergquist, J. *Anal. Chem.* **2002**, *74*, 5826–5830.
- (32) Guo, D. C.; Mant, C. T.; Hodges, R. S. *J. Chromatogr.* **1987**, *386*, 205–222.
- (33) Hodges, R. S.; Parker, J. M.; Mant, C. T.; Sharma, R. R. *J. Chromatogr.* **1988**, *458*, 197–210.
- (34) Bohr, H. G.; Røgen, P.; Jalkanen, K. J. *Comput. Chem.* **2001**, *26*, 65–77.
- (35) Anker, L. S.; Jurs, P. C. *Anal. Chem.* **1992**, *64*, 1157–1164.
- (36) Glen, R. C.; Rose, V. S.; Lindon, J. C.; Ruane, R. J.; Wilson, I. D.; Nicholson, J. K. *J. Planar Chromatogr.* **1991**, *4*, 432.
- (37) Peterson, K. L. *Anal. Chem.* **1992**, *64*, 379–386.
- (38) Yan, A.; Zhang, R.; Liu, M.; Hu, Z.; Hooper, M. A.; Zhao, Z. *Comput. Chem.* **1998**, *5*, 405–412.
- (39) Yan, A.; Jiao, G.; Hu, Z.; Fan, B. T. *Comput. Chem.* **2000**, *24*, 171–179.
- (40) Cupid, B. C.; Nicholson, J. K.; Davis, P.; Ruane, R. J.; Wilson, I. D.; Glen, R. C.; Rose, V. S.; Beddell, C. R.; Lindon, J. C. *Chromatographia* **1993**, *37*, 241–249.
- (41) Sacchero, G.; Bruzzoniti, M. C.; Sarzanini, C.; Mentasti, E.; Metting, H. J.; Coenegracht, M. J. *J. Chromatogr., A* **1998**, *799*, 35–45.
- (42) Madden, J. E.; Avdalovic, N.; Haddad, P. R.; Havel, J. *J. Chromatogr., A* **2001**, *910*, 173–179.
- (43) Lochmüller, C. H.; Reese, C.; Aschman, A. J. *J. Chromatogr., A* **1993**, *656*, 3–18.
- (44) Kaliszcan, R. *Structure and Retention in Chromatography: A Chemometric Approach*; Harwood Academic Publishers: Amsterdam, The Netherlands, 1997.
- (45) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists: an Introduction*; VCH Verlagsgesellschaft: Weinheim, 1993.

predictive capability of any system is strongly dependent on the quality of the acquired retention time data. As a result, an approach based on a genetic algorithm (GA) has been used for the normalization of any potential variabilities of the training retention time data sets. The GA allowed ANN training and testing of its predictive capability using large sets of confidently identified peptides and their retention times for *Deinococcus radiodurans* and *Shewanella oneidensis* microorganisms. The model's predicted retention time information is shown to increase the confidence of peptide identifications.

## EXPERIMENTAL SECTION

**Preparation of Tryptic Digests from *D. radiodurans* and *S. oneidensis*.** *D. radiodurans* and *S. oneidensis* cells were cultured in TGY medium to an approximate 600OD of 1.2 and harvested by centrifugation at 10000g at 4 °C. Prior to lysis, cells were resuspended and washed three times with 100 mM ammonium bicarbonate and 5 mM EDTA (pH 8.4). Cells were lysed by beating with 0.1-mm acid zirconium beads for three 1-min cycles at 5000 rpm. The samples were incubated on ice for 5 min between each cycle of bead-beating. The supernatant containing soluble cytosolic proteins was recovered after centrifugation at 15000g for 15 min to remove cell debris. Proteins were denatured and reduced by addition of guanidine hydrochloride (6 M) and DTT (1 mM), respectively, followed by boiling for 5 min. Prior to digestion, samples were desalted using a 5000 molecular weight cutoff "D-salt" gravity column (Pierce, Rockford, IL) equilibrated in 100 mM ammonium bicarbonate (pH 8.4). Proteins were enzymatically digested at an enzyme/protein ratio of 1:50 (w/w) using sequencing grade modified trypsin (Promega, Madison, WI) at 37 °C for 16 h.

**Capillary LC Coupled with ESI-MS.** HPLC-grade water and acetonitrile were purchased from Aldrich (Milwaukee, WI). Fused-silica capillary columns (30–60 cm, 150- $\mu$ m i.d.  $\times$  360- $\mu$ m o.d., Polymicro Technologies, Phoenix, AZ) packed with 5- $\mu$ m C18 particles were manufactured in-house as described previously.<sup>12</sup> Briefly, capillary RPLC was performed using an ISCO LC system (model 100DM, ISCO, Lincoln, NE). The mobile phases for gradient elution were (A) acetic acid/TFA/water (0.2:0.05:100 v/v) and (B) TFA/acetonitrile/water (0.1:90:10, v/v). The mobile phases, delivered at 5000 psi using two ISCO pumps, were mixed in a stainless steel mixer (~2.8 mL) with a magnetic stirrer before flow-splitting and entering the separation capillary. In this way, a nonlinear (exponential) gradient is generated, as has been previously described,<sup>46</sup> providing an analysis time of ~180 min. Fused-silica capillary flow splitters (30- $\mu$ m i.d. with various lengths) were used to manipulate the gradient speed. Capillary RPLC was coupled on-line with MS through an ESI interface (a stainless steel union was used to connect an ESI emitter and the capillary separation column).

The peptide database has been generated by using several mass spectrometers, including 3.5-, 7-, and 11.4-T FTICR instruments (described in detail previously<sup>47</sup> and references therein),

as well as several ion-trap mass spectrometers (LCQ, LCQ Duo, LCQ DecaXP; ThermoFinnigan, San Jose, CA). The ANN software used was NeuroWindows version 4.5 (Ward Systems Group, USA) and utilized a standard back-propagation algorithm on a Pentium 1.5 GHz personal computer.

## RESULTS AND DISCUSSION

**Brief Description of the Artificial Neural Network.** In comparison with classical statistical methods, ANN-based approaches have advantages that include a capacity to self-learn and to model complex data without the need for a detailed understanding of the underlying phenomena.

A feed-forward neural network model, sometimes called a back-propagation neural network due to its most common learning algorithm, is used in this work. It is composed of a large number of neurons, nodes, or processing elements organized into a sequence of layers.<sup>48,49</sup> The architectures of these ANN models contain at least two layers, an input layer with one node for each variable in a data vector and an output layer consisting of one node for each variable to be investigated. Additionally, one or more hidden layers can be added between the input and output layers if the complexity of the data so require. Nodes in any layer can be fully or partially connected to nodes of a succeeding layer, as shown in Figure 1, where each hidden or output node receives signals in parallel. The input signal to a node is modulated by a weight ( $w$ ) along each link. The net input to a node is thus a function of all signals to a node and all of its associated weights. For example, the net input for a node  $j$  is given by

$$\text{net}_j = \sum_i w_{ji} O_i \quad (1)$$

where  $i$  represents nodes in the previous layer,  $w_{ji}$  is the weight associated with the connection from node  $i$  to node  $j$ , and  $O_i$  is the output of node  $i$ .

The final output signal of a node is usually confined to a specified interval, say between 0 and 1. We are forced, then, to make the net input to the neuron undergo an additional transformation through or using a transfer function. There are several transfer functions available, satisfying a requirement of continuity, set by the back-propagation algorithm. The most popular one is the sigmoid function given by

$$O_j = \frac{1}{(1 + e^{-\text{net}_j})} \quad (2)$$

In essence, these equations applied to nodes in the hidden and output layers allow these ANNs to perform multivariate nonlinear regression using a sigmoidal function, and because of the parallel processing of nodes within each layer, these ANNs have the ability to learn multivariate nonlinear functions.

The process of adapting the weights to an optimum set of values is called training the neural network. To train the neural network, there exist several training algorithms, of which the back-propagation algorithm is one.<sup>50</sup>

(46) Shen, Y.; Tolić, N.; Zhao, R.; Paša-Tolić, L.; Li, L.; Berger, S. J.; Harkewicz, R.; Anderson, G. A.; Belov, M. E.; Smith, R. D. *Anal. Chem.* **2001**, *73*, 3011–3021.

(47) Harkewicz, R.; Belov, M. E.; Anderson, G. A.; Paša-Tolić, L.; Masselon, C. D.; Prior, D. C.; Udseth, H. R.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 144–154.

(48) Werbos, P. J. Ph.D. Thesis, Harvard University, Cambridge, MA, 1974.

(49) Werbos, P. J. *The Roots of Back-Propagation*; John Wiley & Sons: New York, 1994.



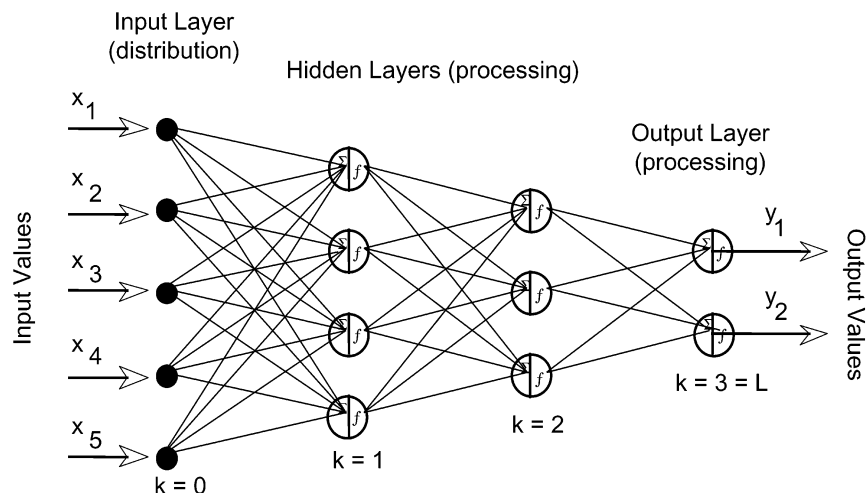


Figure 1. Typical three-layer neural network showing the flow of signals from left to right.

**Normalization of Peptide Elution Times Using a Genetic Algorithm.** An “intelligent” algorithm for the normalization of retention time was desired to compare a large number of LC/MS experiments, because of the variability associated with constant high-pressure capillary LC separations using syringe pumps.<sup>11,51,52</sup> Small changes in split ratio, column lengths, column packings, void volumes, etc. unavoidably lead to some retention time variability. Thus, all peptide retention times were normalized to the range [0, 1] by using a genetic algorithm (GA).

A GA is an algorithm based on evolutionary computation and survival of the fittest and is often applied to optimization problems, such as optimizing the free variables in a hypothesis function.<sup>53,54</sup> Solutions to problems are coded as individuals, which evolve through generations. An individual in our coding is a vector of real values (line functions) of slopes and intercepts for each experiment being normalized. The fittest individuals in each generation breed the next generation through crossover and mutation operators, recombining best “genes.” The “genes” in the offspring are then perturbed in the next step by a small value and a small probability. This iterative process causes the best solution, the fittest individuals, to be incrementally refined.

The GA was applied to 51 150 (9121 different) peptides identified from 687 LC/MS/MS analyses to establish a common timeline so that the same peptides’ variances of normalized elution times (NETs) across the different separations were minimized. The GA was set up to optimize the two linear equation variables,  $k$  and  $m$  in  $y = kx + m$ , for each experiment (Figure 2 shows coding of the individuals): one variable ( $m$ ) normalized the start of the recording time and the other ( $k$ ) normalized the gradient speed. The GA optimized these two variables for each separation to reduce the variance function of specific peptides, i.e., the

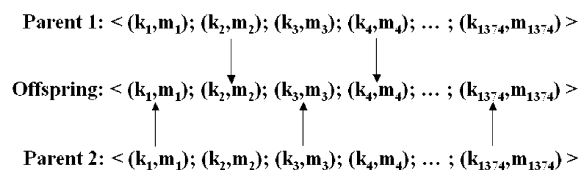


Figure 2. Recombination of two parents’ genes into a new offspring in a genetic algorithm.

regressed elution times for each separation. This optimization scheme of multiple linear regressions normalized the peptide elution times into a common [0, 1] range.

The normalization of elution times to a 0–1 range is based on six peptides identified frequently in both *D. radiodurans* and *S. oneidensis*. The peptides were the following: (1) YNQLLR, (2) IVSLAPEVL, (3) VPLHTLR, (4) TFAIPHGGGGPGMGPIGVK, (5) ELATAK, and (6) PGVVIGK. All experiments were first normalized with the GA to minimize elution time variances for the same peptides across the set of all peptides. With all peptide elution times on the same scale, the means for the peptide elution times were regressed/fitted to a 0.1–0.9 scale of normalized elution times. The 0.1–0.9 range was chosen for the shortest and longest retained peptides, respectively, rather than 0–1, to accommodate for actual peptides eluting before the mean of 0.1 and later than the mean of 0.9 and to accommodate for peptides not yet identified that may elute earlier or later than in the set of identified peptides. These six identified peptides with their normalized elution times served as calibrant peptides in our GA when normalizing the *D. radiodurans* and *S. oneidensis* datasets. However, individual experiments are not regressed against the elution times for these peptides; rather, the means for these six peptides in the whole species database were loosely regressed against the calibrant peptide values while the GA was also minimized for the NET variance of all peptides in the database for that species.

The average variance of NETs for the peptides identified in more than one experiment is 0.000 276 (standard deviation 0.016 615), or the average normalized elution time deviates from its mean by 1%. In future studies, a set of standard peptides will be selected that elute at the beginning, the middle, and the end of the chromatogram to further improve the normalization of retention times.

- (50) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. *Learning internal representations by error propagation*, *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, Rumelhart, D. E.; McClelland, J. L.; (Eds.), MIT Press: Cambridge, MA, 1986; Vol. 1, Foundations; pp 318–362.
- (51) Macnair, J. E.; Lewis, K. C.; Jorgenson, J. W. *Anal. Chem.* **1997**, *69*, 983–989.
- (52) Macnair, J. E.; Patel, K. D.; Jorgenson, J. W. *Anal. Chem.* **1999**, *71*, 700–708.
- (53) Holland, J. H. *Adaptation in Natural and Artificial Systems*, University of Michigan Press: Ann Arbor, MI, 1975.
- (54) Goldberg, D. E. *Genetic Algorithms in Search, Optimisation and Machine Learning*, Addison-Wesley: Reading, MS, 1989.

Table 1. Mean Square Errors (MSEs) as a Function of the Number of Hidden Layer Nodes in Seven Training Sessions

hidden layer nodes	training data MSE	cross-validation MSE
0	0.01	0.007 2
2	0.009 45	0.007 02
3	0.009 2	0.006 88
4	0.009 02	0.006 95
5	0.008 84	0.006 84
15	0.008 44	0.006 85
30	0.008 7	0.006 9

**The Use of Artificial Neural Networks for Peptide Elution Time Prediction.** The ANN training set consisted of 6958 confidently identified *D. radiodurans* peptides measured by the RPLC/ESI-MS/MS and further verified with high-mass measurement accuracy using RPLC/ESI-FTICR-MS to exist in the *D. radiodurans* polypeptide mixture.<sup>8,15</sup> Each peptide was coded as a 20-dimensional vector consisting of the normalized number of each of the 20 amino acid residues making up the peptides. Each residue count was normalized to a fraction of the maximum count of that residue in any peptide in the *D. radiodurans* database. These peptide code vectors were repeatedly input into the ANN by the back-propagation algorithm to reduce output error. The output error is the squared difference between a target value of the ANN and the predicted value. In this case, the target values were the known NETs of the peptides. The ANN thus learned the relationship between the coded peptide vectors and their measured NETs.

The hidden layer(s) configuration for the ANN was empirically determined by using a cross-validation data set during training. In general, a hidden layer with too few nodes may not sufficiently model the data. A hidden layer with too many nodes may overfit the data in the training set and not provide an effective predictive capability for new data. The ANN was trained with 97% of the DR peptides and cross-validated with the remaining data. Typically, the cross-validation data sets are used to stop the training when the error for the data set ceases to decrease. Going beyond this point suggests that the ANN “learns” from noise in the training set that is not present in the cross-validation set. Our experience in training ANNs with peptide elution data was that the ANN could not be over-trained. Both the errors on the training and the cross-validation data sets rapidly converged to minimum values. A small improvement was realized by using a two-node hidden layer instead of no hidden layers. Increasing to three hidden nodes made an even smaller improvement. Table 1 shows error rates as a function of the number of hidden layer nodes in seven training sessions. The hidden layer could be increased to a large number of nodes without the back-propagation algorithm being able to reduce the errors or being able to overfit the data. We used a hidden layer with two nodes for the presented work, since it reduced the error to a near optimal level without potentially sacrificing generality. The training was stopped at 1000 epochs, because the errors appeared to have converged at different learning rates ranging from 0.001 to 0.1. The final ANN model with 20 input, 2 hidden, and 1 output nodes (20–2–1) is depicted in Figure 3.

Table 2 summarizes the calculated ANN weights of amino acid residues after training with the *D. radiodurans* peptides. From

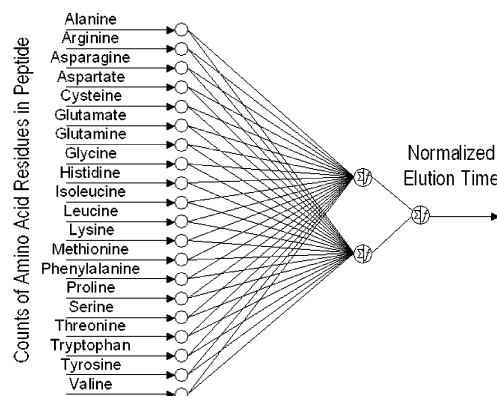


Figure 3. The 20–2–1 neural network architecture used in this study.

Table 2. Artificial Neural Network Weights of Amino Acid Residues; Comparison with Previous Retention Coefficients (RC) for Amino Acid Residues

amino acid	ANN weight	RC			
		Meek et al. <sup>20</sup>	Casa et al. <sup>30</sup>	Guo et al. <sup>22</sup>	Browne et al. <sup>21</sup>
leucine	6.12	11	10.47	8.1	20
phenylalanine	3.37	13.4	16.31	8.1	19.2
isoleucine	2.37	8.5	7.76	7.4	6.6
tryptophan	2.27	17.1	18.65	8.8	16.3
methionine	1.63	5.4	8.31	5.5	5.6
valine	1.63	5.9	6.42	5	3.5
tyrosine	0.72	7.4	7.37	4.5	5.6
alanine	0.71	1.1	−0.67	2	7.3
glutamate	0.56	0.7	0.08	1.1	−7.1
proline	0.48	4.4	2.19	2	5.1
cysteine	0.32	7.1	N.I.	2.6	−9.2
aspartate	0.18	−1.6	0.77	0.2	−2.9
threonine	0.18	−1.7	4.21	0.6	0.8
glycine	−0.21	−0.2	−0.25	−0.2	−1.2
arginine	−0.24	−0.4	−0.01	−0.6	−3.6
asparagine	−0.29	−4.2	N.I.	−0.6	−5.7
glutamine	−0.3	−2.9	N.I.	0	−0.3
serine	−0.35	−3.2	−2.18	−0.2	−4.1
lysine	−0.55	−1.9	−3.2	−2.1	−3.7
histidine	−0.59	−0.7	−2.99	−2.1	−2.1

the weights, we see that leucine is the amino acid that most affects peptide retention times. In comparison with previous work, only Browne et al.<sup>21</sup> measured Leu as the amino acid residue having the highest retention coefficient. Generally, our results are in good agreement with those of Guo et al.<sup>22</sup> (same positive and negative sign as well as a similar ordering of amino acid residue dependence upon retention time). The similarity of the chromatographic conditions, C18 columns, TFA/water/acetonitrile-based gradient elution are the obvious reasons for the similarities.

The ANN model was evaluated using peptides identified from the microorganism *S. oneidensis* using RPLC/ESI-ion trap MS/MS. The average error for predicting *S. oneidensis* NETs for 7080 peptides from 157 analyses was 0.047983 or ~4.8%. Figure 4a shows a plot of the predicted NETs for 7080 *S. oneidensis* peptides identified in 157 different separations. These results should be considered worst case because of the uncertainty in peptide identifications; *S. oneidensis* peptides, unlike the *D. radiodurans* peptides in the training set, were not validated using accurate mass measurements. Furthermore, the data in Figure 4a suggests the extremes in errors for LC elution predictions, but does not clearly

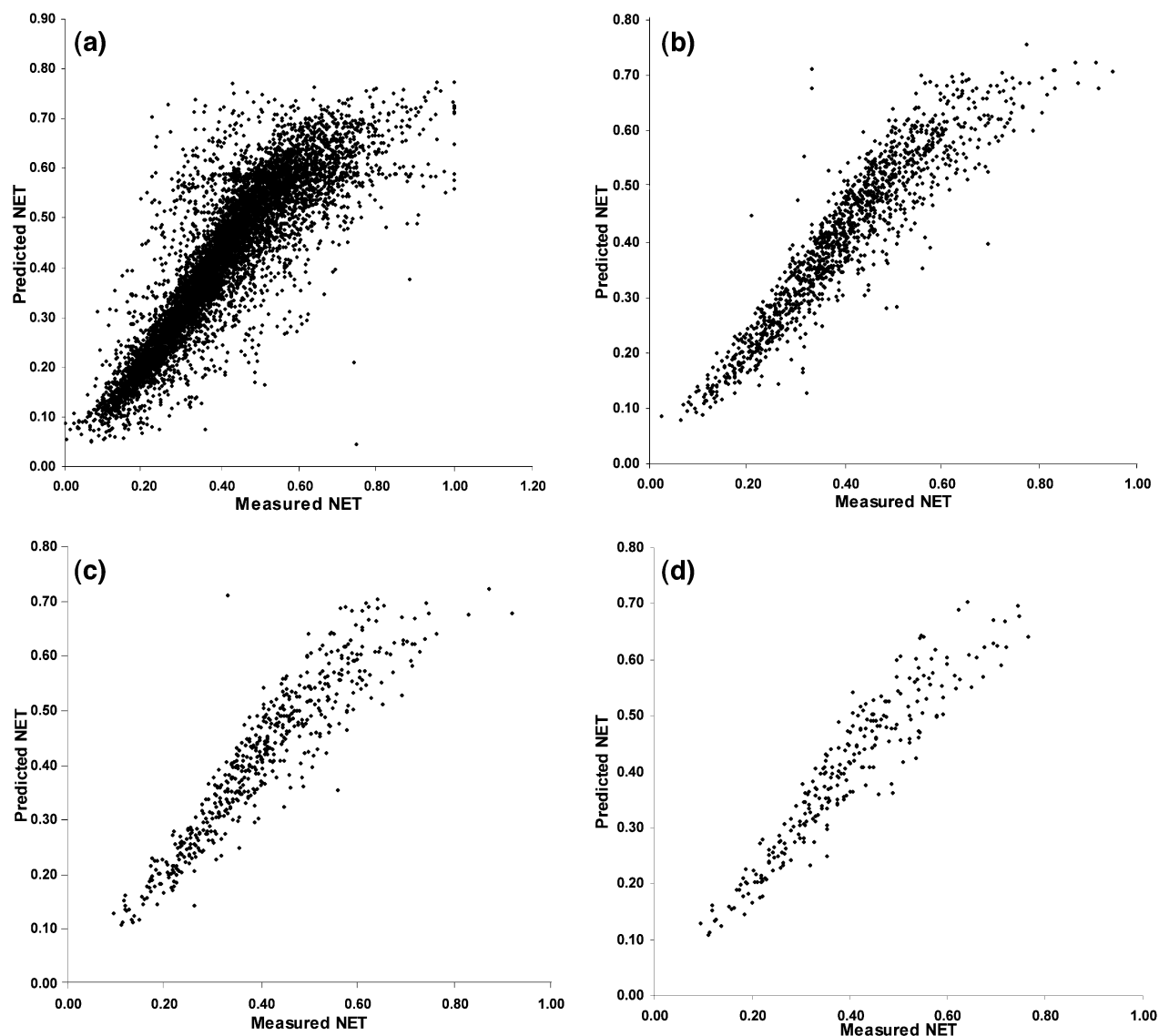


Figure 4. Measured vs predicted normalized elution times among 42 378 *S. oneidensis* peptides from 157 experiments which have been identified at least (a) 3 times (7080 peptides), (b) 20 times (1270 peptides), (c) 40 times (536 peptides), and (d) 60 times (259 peptides).

show the distribution of errors around the mean. The plot also includes errors due to variations between separations, which are not fully eliminated in the normalization process (and would benefit from the use of elution time calibrants).

A significant number of *S. oneidensis* peptides were identified one or only a few times across all 157 experiments, suggesting that they may be misidentifications. This is supported by the observation that the average prediction error decreases rapidly when the model is tested with peptides required to occur in an increasing number of experiments. A more rigorous error measurement from *S. oneidensis* peptides from the same number of experiments and each peptide occurring at least 20, 40, or 60 times to reduce spurious misidentifications yielded an average error of 3.86, 3.67, and 3.66%, respectively, (see Figure 4b–d). It can be seen that the peptides with poor correlation with our model (i.e., highly dispersed in the plots) are eliminated when only the peptides that occurred at least 60 times were selected, again suggesting that infrequently seen peptides are possibly misidentified. Furthermore, our preliminary LC–FTICR experiments of the *S. oneidensis* imply that peptides with multiple ion-trap

identifications are probably correct based on accurate mass measurements. Thus, as the probability of correct identifications is increased, a better correspondence with predicted elution times is observed.

Figure 5 shows the error distribution of these 1270 *S. oneidensis* peptides that have been identified at least 20 times. This curve is assumed to approach the true distribution of the prediction model's performance for correctly identified peptides. For this peptide set, 50% are predicted within  $\pm 2.97\%$  of the measured NETs, and more than 95% are predicted within  $\pm 10\%$  of the measured NETs.

One of the major advantages of our model in relation to previous ones is that it provides more accurate prediction for longer peptides. As can be seen from Table 3, the average error is very low for peptides up to 20-mer size; the error then increases just slightly for longer peptides.

The very fact that not all peptides can be correctly identified by either accurate mass measurement or MS/MS experiments has prompted this research into utilizing elution time as an additional metric for identifying peptides. The use of peptide

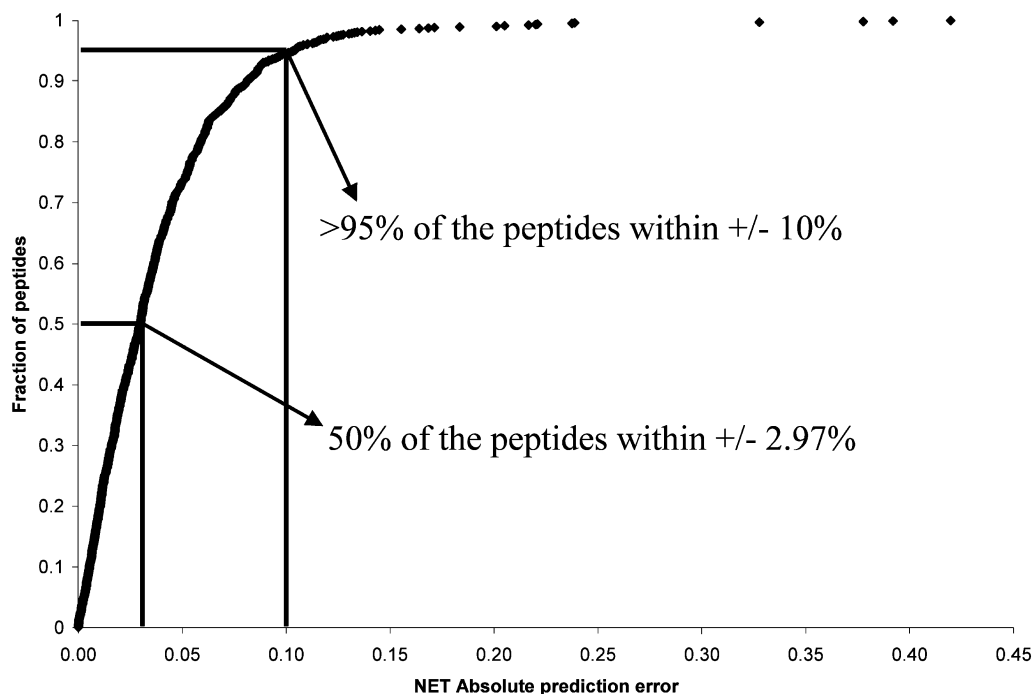


Figure 5. Prediction error distribution for 1270 *S. oneidensis* peptides that were tentatively identified at least 20 times using the program SEQUEST. The graph shows the fraction of peptides vs the NET error levels. For example, the graph shows that 50% of the peptides have <3% prediction error, and more than 95% have <10% error.

Table 3. Average Mean Square Error (MSE) of the Peptide Elution Time Prediction in Relation to the Peptide Length<sup>a</sup>

peptide length	peptides with that length	av MSE
5–10	353	0.001 73
11–20	618	0.003 06
21–30	258	0.005 60
31–40	36	0.006 13
41–	5	0.007 66

<sup>a</sup> These are 1270 *S. oneidensis* peptides identified at least 20 times (same peptides as in Figure 4b).

elution prediction will be particularly interesting for the identification of isobaric peptides by LC–FTICR. As can be seen from Figure 6, it was possible to distinguish between the isobaric *D. radiodurans* peptides LPNHIQVDDLRLQQLDV and VAINDTD-NHTLAHLLK as a result of their significantly different elution times, accurately predicted with our model. Although these two peptides have the same molecular formula, interestingly, they have different charges. Furthermore, as shown from Table 4, several isobaric peptides (undistinguishable even with 1 ppm mass accuracy) have different retention times and were identified with our model. Moreover, it is also possible to distinguish isomeric peptides, which have different Ile/Leu ratios (i.e., IVIEIK and VILLEK) due to the different ANN weights assigned to these amino acid residues. Some of the peptides, of course, will have very similar retention times (i.e., the isobar peptides ANAAINS-GAFK and IIAAGANVVR have the same NET = 0.26, data not shown). This approach will be even more useful for proteomes of higher complexity, in which the number of possible peptides is greatly increased. For example, in a typical 7 ppm “window” between 1605.851 and 1605.863 Da, the human proteome codes

for 12 tryptic peptides, but three peptides (QTFEAAILTQLHPR, TLHSLTQWNGLINK, and LLFLVGTASNPHPEAR) have masses of 1605.862 64 Da and are indistinguishable by mass. Importantly, however, these peptides are predicted to have different predicted LC retention times.

It must be pointed out here that this model applies for the present set of experimental conditions. In the case of separations using different stationary phases, mobile phases, temperatures, etc., the present system would have to be “recalibrated” or properly “mapped” onto the new set of conditions, and perhaps a new training data set would need to be developed for the generation of new ANN weights.

Furthermore, the model described takes into account only the peptides’ amino acid composition, not their sequence. Isomeric peptides (same amino acids in a different order) are predicted to elute at the same time, although it has been shown that such peptides are often separated in LC. Moreover, sequence-dependent effects, such as conformational and nearest-neighbor effects, may be additional factors for deviations from predicted retention times. Indeed, recently, Wimley et al.<sup>55</sup> showed that occlusion effects may occur in the case of guest (X) side chains in the host–guest pentapeptides ACWL-X-LL that may lead in changes in the overall hydrophobicity of the peptide. Another sequence-dependent effect that leads to “anomalous” retention times in liquid chromatography is due to conformational differences (i.e., helical vs not helical peptides and amphipathic helical vs nonamphipathic helical).<sup>56–60</sup>

(55) Wimely, W. C.; Creamer, T. P.; White, S. H. *Biochemistry* **1996**, *35*, 5109–5124.

(56) Houghten, R. A.; Degraw, S. T. *J. Chromatogr.* **1987**, *386*, 223–228.

(57) Büttner, K.; Pinilla, C.; Appel, J. R.; Houghten, R. A. *J. Chromatogr.* **1992**, *625*, 191–198.

(58) Sereda, T. J.; Mant, C. T.; Sönnichsen, F. D.; Hodges, R. S. *J. Chromatogr.* **1994**, *676*, 139–153.

VAINDLTDNHTLAHLLK

$C_{83}H_{138}N_{24}O_{26}$

$[M+H] = 1887.02$

Charge = +2

PredNET = 0.521

$m/z$  944.4851-944.5523

NET = 0.521

EIC

LPNHIQVDDLRLQLLDV

$C_{83}H_{138}N_{24}O_{26}$

$[M+H] = 1887.02$

Charge = +3

PredNET = 0.333

$m/z$  630 – 630.1

NET = 0.332

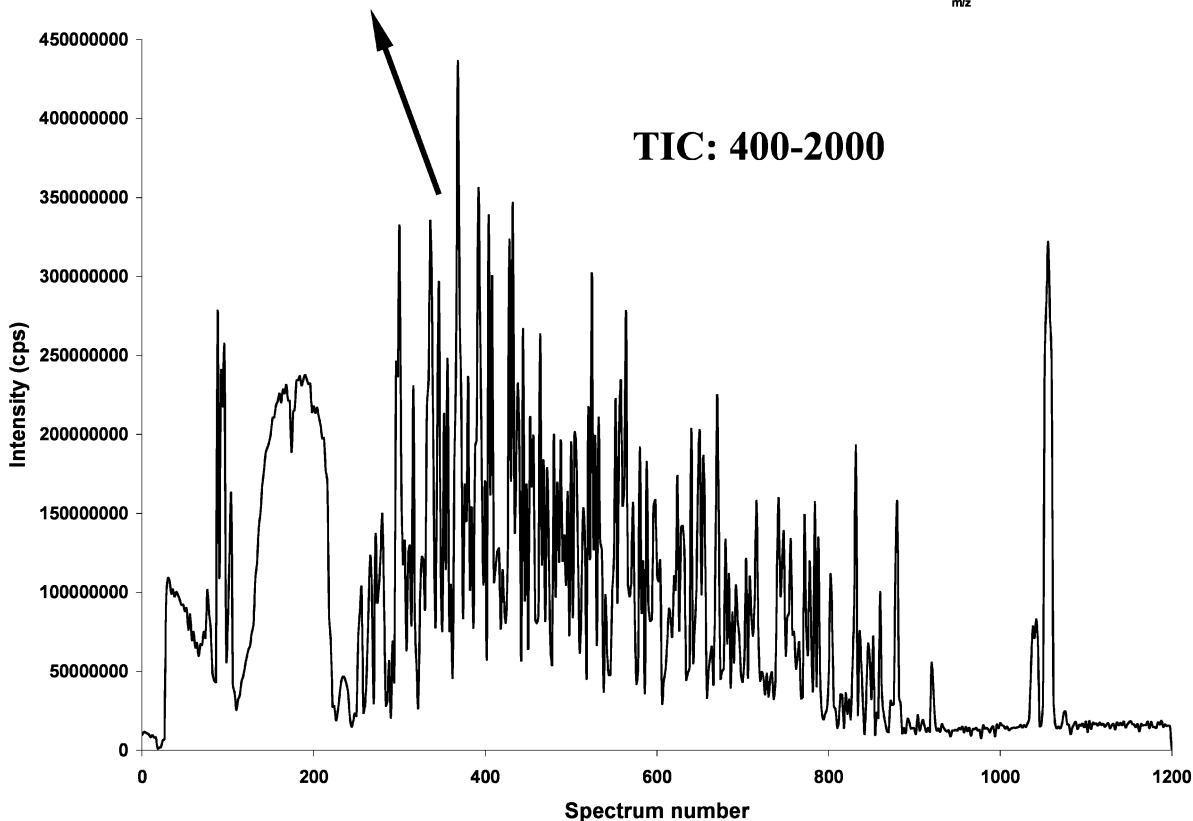


Figure 6. Example of the more confident identification of two isobaric peptides by using peptide predicted elution times as an additional metric. The isobaric peptides LPNHIQVDDLRLQLLDV and VAINDLTDNHTLAHLLK have a different normalized elution time (NET), which has allowed their differentiation. The figure shows the total ion current (TIC), the corresponding extracted ion currents (EIC), and the mass spectra of these peptides.

In this regard, Zhou et al.<sup>61</sup> showed that peptides with the same amino acid composition have different  $\alpha$ -helical contents. The

more amphipathic the peptide, the higher the helicity. Amphipathic  $\alpha$ -helical peptide conformations have been shown to have

(59) Purcell, A. W.; Aguilar, M. I.; Wettenhall, R. E. H.; Hearn, M. T. W. *Pept. Res.* **1995**, *8*, 160–170.

(60) Sereda, T. J.; Mant, C. T.; Hodges, R. S. *J. Chromatogr., A* **1995**, *695*, 205–221.



Table 4. Actual and Predicted Normalized Elution Time (NET) Values of Several Isobaric Peptides from Both Microorganisms, *D. radiodurans* (D.R.) and *S. oneidensis* (S.O.)

microorganism	peptide	MW	NET		Abs. error
			predicted	actual	
D.R.	GVNIR	556.3285	0.4186	0.4164	0.00215
	IAQAR	556.3285	0.1529	0.1349	0.01796
S.O.	IAGLLR	640.4271	0.3250	0.3351	0.01016
	VIAAIR	640.4271	0.2710	0.2589	0.01217
S.O.	AAIEAAK	671.3853	0.1318	0.1474	0.01568
	DALLNK	671.3853	0.2256	0.2390	0.01334
S.O.	IVIEIK	712.4734	0.3230	0.3243	0.00123
	VILLEK	712.4734	0.3332	0.3311	0.00214
D.R.	DKETLPR	856.4607	0.4506	0.4249	0.02570
	IAEQIER	856.4607	0.2032	0.2013	0.00191
D.R.	ERAQALLR	954.5563	0.4962	0.4630	0.03317
	RVGQDLIR	954.5563	0.2702	0.2655	0.00473
S.O.	DLSVEELR	958.4971	0.3134	0.3376	0.02420
	EAVDGDVKV	958.4971	0.1447	0.1610	0.01627
S.O.	SGNEFNVGSLVFR	1423.7090	0.4668	0.4754	0.00852
	YGFDIRPASNAK	1423.7090	0.3329	0.3323	0.00058
D.R.	LPNHIQVDDLRLQLLDV	1886.0214	0.3330	0.3322	0.00077
	VAINDLTDNHTLAHLK	1886.0214	0.5206	0.5213	0.00073
S.O.	AIPQSVGEQSIPSLAPMLER	2122.1140	0.4805	0.4601	0.02048
	PEAAVMIQADKDTTHGLVVK	2122.1140	0.4173	0.4486	0.03131
D.R.	AITVLSALSGILMAQTPAWQIISPPELSVMAGAGIGALAG	3931.1155	0.5481	0.5201	0.02797
	SSPLFSTQLALALAVRLCLLTPAEALSACTVNAAAYALGL	3931.1155	0.4611	0.4544	0.00669

increased retention times relative to their elution times calculated using retention coefficients.<sup>56–61</sup> This is because amphipathic peptides interact with the hydrophobic phase in such a way that the hydrophobic part of the structure strongly interacts with the reversed stationary phase while the polar groups remain in contact with the hydrophilic mobile phase.<sup>60–62</sup> Until now, it has not been investigated if in practice peptides obtained from tryptic digestions yield amphipathic  $\alpha$  helices or not. Finally, it has been recently shown that even very small isomeric peptides may elute in different retention times, implying that the peptide structure might not be the only parameter governing their LC retention time.<sup>63</sup> For example, in a 28-min LC run, Gly-Leu and Leu-Gly were eluted with a 3.5-min difference.<sup>63</sup>

Clearly, the development of more sophisticated ANNs incorporating selected sequence features offer the likelihood of further improvements in predictions, including the ability to distinguish sequence variations. The problem is that larger experimental datasets relative to different peptide retention times should be available to include some aspects of sequence information in the ANN. Finally, our initial model has an inherent weakness relative to the use of tryptic peptides for its training. Thus, the peptides used in this study include Arg and Lys only once except when missed cleavages occur. Because of their basic character, these amino acids change the  $pK_a$ /apparent charge of these peptides and, consequently, their retention times. As a result, the values given for Arg and Lys might not apply for nontryptic peptides having additional Arg or Lys residues in their structure. While

this should not be a problem in the case of ideal trypsin proteolysis, such missed cleavages are commonly observed in global proteomic studies. In future work, the ANN will be trained to more correctly predict retention times for peptides containing more than one Lys or Arg residue.

## CONCLUSIONS

Artificial neural networks have been developed and demonstrated for the prediction of tryptic peptide elution times. An ANN has been trained with the experimental results derived from the proteome of the microorganism *D. radiodurans* and was then used for the successful elution time prediction for tryptic peptides from *S. oneidensis*. The use of different species for the training and testing of the ANN demonstrate the unbiased nature of this method. Despite the simplicity of our initial model (only the amino acid residue counts are taken under consideration), the average accuracy achieved was  $\sim 3\%$ .

The capability for elution time prediction of peptides adds another dimension of information for proteomic efforts, because it either allows new peptide identifications to be made or increases the confidence of the peptide identifications. Such capabilities will be particularly useful in isobaric peptide identifications in conjunction with accurate mass information. Future development of this method will aim to increase the accuracy of our ANN model by including more information related to amino acid sequence (particularly factors that influence secondary structure) and is expected to improve predictions and provide new capabilities (e.g., to predict the retention time of isomeric amino acids). We also aim to improve the normalization process by using a set of (calibrant) standard peptides in each run and by further increasing both the size and quality (i.e., confidence of identification for the training set). Furthermore, we plan to explore the use of ANNs

(61) Zhou, N. E.; Mant, C. T.; Hodges, R. S. *Pept. Res.* **1990**, *3*, 8–20.

(62) Wieprecht, T.; Rothmund, S.; Bienert, M.; Krause, E. *J. Chromatogr., A* **2001**, *912*, 1–12.

(63) Petritis, K.; Brusaux, S.; Guenu, S.; Elfakir, C.; Dreux, M. *J. Chromatogr., A* **2002**, *957*, 173–185.

to predict the elution time of peptides with posttranslational modifications. Finally, the ANN approach will be applied to the elution time prediction of peptides separated by ion-exchange chromatography.

#### ACKNOWLEDGMENT

This work has been supported by the United States Department of Energy Office of Biological and Environmental Research,

Life Sciences Division. Pacific Northwest National Laboratory is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract DE-AC06-76RLO 1830.

Received for review August 7, 2002. Accepted December 31, 2002.

AC0205154