Anal. Chem. **2006,** *78,* 5026−5039

Improved Peptide Elution Time Prediction for Reversed-Phase Liquid Chromatography-MS by Incorporating Peptide Sequence Information

Konstantinos Petritis,[†] Lars J. Kangas,[‡] Bo Yan,[§] Matthew E. Monroe,[†] Eric F. Strittmatter,[†] Wei-Jun Qian,[†] Joshua N. Adkins,[†] Ronald J. Moore,[||] Ying Xu,[§] Mary S. Lipton,[†] David G. Camp, II,[†] and Richard D. Smith*[,†]

Biological Sciences Division, Computational Sciences and Mathematics Division, and Environmental and Molecular Sciences Laboratory, Pacific Northwest National Laboratory, P. O. Box 999, Richland, Washington 99352, and Computational System Biology Laboratory, Biochemistry and Molecular Biology Department, University of Georgia, Athens, Georgia 30601

We describe an improved artificial neural network (ANN)-based method for predicting peptide retention times in reversed-phase liquid chromatography. In addition to the peptide amino acid composition, this study investigated several other peptide descriptors to improve the predictive capability, such as peptide length, sequence, hydrophobicity and hydrophobic moment, and nearest-neighbor amino acid, as well as peptide predicted structural configurations (i.e., helix, sheet, coil). An ANN architecture that consisted of 1052 input nodes, 24 hidden nodes, and 1 output node was used to fully consider the amino acid residue sequence in each peptide. The network was trained using ~345 000 nonredundant peptides identified from a total of 12 059 LC−MS/MS analyses of more than 20 different organisms, and the predictive capability of the model was tested using 1303 confidently identified peptides that were not included in the training set. The model demonstrated an average elution time precision of ~1.5% and was able to distinguish among isomeric peptides based upon the inclusion of peptide sequence information. The prediction power represents a significant improvement over our earlier report (Petritis, K.; Kangas, L. J.; Ferguson, P. L.; Anderson, G. A.; Pasa-Tolic, L.; Lipton, M. S.; Auberry, K. J.; Strittmatter, E. F.; Shen, Y.; Zhao, R.; Smith, R. D. *Anal. Chem.* 2003, *75*, 1039−1048) and other previously reported models.

The analysis of peptides (e.g., tryptically digested proteins) by on-line coupling of liquid chromatography (LC) with electrospray ionization-mass spectrometry (ESI-MS) is presently the most common approach for characterizing complex proteomes. While several methods and software tools are available for identifying peptides/proteins from mass spectra, the high complexity of a digested proteome (containing thousands or even millions of detectable peptides) and the vastly larger number of possible peptide sequences make accurate peptide/protein identification challenging; final results can include large numbers of false positive identifications.[1]

We have been working to apply additional information such as LC retention time to improve the confidence in peptide identifications, an approach also recently suggested by a group working to establish publication guidelines for peptide and protein identification.[2] The use of peptide retention time information has proved useful in the past for LC method development of simple peptide mixtures,[3,4] purification of peptides of interest,[4,5] and identification of simple peptide mixtures in conjunction with UV and/or fluorescence and/or colorimetric methods of detection.[6]

Efforts to predict the chromatographic behavior of peptides on the basis of amino acid composition are not new. In 1951, Knight[7] and Pardee[8] showed that synthetic peptide retardation factors (R_f) in paper chromatography could be predicted with some accuracy. In 1952, Sanger[9] demonstrated that peptides of the same amino acid composition, but different sequence, could frequently be separated. More recently, there have been several reports on the prediction of peptide elution times in reversed-phase (RP)[10−15]

* To whom correspondence should be addressed. E-mail: rds@pnl.gov.

[†] Biological Sciences Division, Pacific Northwest National Laboratory.

[‡] Computational Sciences and Mathematics Division, Pacific Northwest National Laboratory.

[§] University of Georgia.

[||] Environmental and Molecular Sciences Laboratory, Pacific Northwest National Laboratory.

(1) Cottingham, K. *Anal. Chem.* **2004**, *76*, 95A−97A.
(2) Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskii, A. *Mol. Cell. Proteomics* **2004**, *3*, 531−533.
(3) Sanz-Nebot, V.; Toro, I.; Barbosa, J. *J. Chromatogr., A* **1999**, *25−38*, 25−38.
(4) Sanz-Nebot, V.; F. Benavente; I. Toro; Barbosa, J. *Anal. Bioanal. Chem.* **2003**, *377*, 306−315.
(5) Browne, C. A.; Bennett, H. P. J.; Solomon, S. *Anal. Biochem.* **1982**, *124*, 201−208.
(6) Perrin, E.; Miclo, L.; Driou, A.; Linden, G. *Anal. Commun.* **1996**, *33*, 143−147.
(7) Knight, C. A. *J. Biol. Chem.* **1951**, *190*, 753−756.
(8) Pardee, A. B. *J. Biol. Chem.* **1951**, *190*, 757−762.
(9) Sanger, F. *Adv. Protein Chem.* **1952**, *7*, 1−7.
(10) Meek, J. L. *Proc. Natl. Acad. Sci. U.S.A.* **1980**, *77*, 1632−1636.
(11) Meek, J. L.; Rossetti, Z. L. *J. Chromatogr.* **1981**, *211*, 15−28.
(12) Guo, D.; Mant, C. T.; Taneja, A. K.; Parker, J. M. R.; Hodges, R. S. *J. Chromatogr., A* **1986**, *359*, 499−517.
(13) Mant, C. T.; Burke, T. W. L.; Black, J. A.; Hodges, R. S. *J. Chromatogr., A* **1988**, *458*, 193−205.
(14) Wilce, M. C. J.; Aguilar, M. I.; Hearn, M. T. W. *J. Chromatogr.* **1991**, *536*, 165−183.
(15) Wilce, M. C. J.; Aguilar, M. I.; Hearn, M. T. W. *J. Chromatogr.* **1993**, *632*, 11−18.

10.1021/ac060143p CCC: $33.50
Published on Web 06/08/2006

and normal-phase[16,17] LC. Most of this reported work used the so-called "retention coefficient" approach, which is based on the summation of empirically determined amino acid residue retention coefficients. The assumption that the chromatographic behavior of peptides is mainly or solely dependent on amino acid composition holds up fairly well for small peptides (up to 15−20 residues), but is inadequate for proteomic applications, e.g., involving tryptic peptides, where the practical upper limit can exceed 50 amino acid residues. Furthermore, with the retention coefficient approach, isomeric peptides are predicted to elute at the same time, which is not the case.[18−21]

Improving peptide retention time prediction in RPLC requires an understanding of the various factors affecting peptide retention. These factors have been thoroughly investigated, and it is now widely accepted that retention behavior of peptides in RPLC is governed by (1) amino acid composition,[10−15] (2) peptide length (or mass),[13,22−24] and (3) sequence-dependent effects.[25−37] The third category can be further divided into nearest-neighbor and conformation effects, where the former is defined to be amino acid sequence-dependent, but independent of peptide conformation.[25] Mant et al.[13] tried to improve peptide retention time prediction by extending the retention coefficient approach by including the peptide length. Krokhin et al.[38] used separate retention coefficients for amino acids at the N-terminus of the peptide in addition to the peptide length, further improving the retention coefficient model. Liu et al.[39] applied a support vector machine and the heuristic method to develop nonlinear and linear models between the capacity factor (log *k*) and seven peptide molecular constitutional and topological descriptors (i.e., number of single bonds, number of rings, etc.), but did not take into account peptide structure. Recently, Kaliszan and co-workers[40,41] used quantitative structure−retention relationships (QSRR) to predict peptide retention times. Descriptors used to derive the necessary QSRR included the logarithm of the sum of retention times of the amino acids that composed the peptide, the logarithm of the van der Waals volume of the peptide, and the logarithm of the peptide calculated 1-octanol−water partition coefficient. Makrodimitris et al.[42] used a mesoscopic simulation that employed Langevin dipoles on a lattice for the solvent and calculated partial charges for the solute to estimate free energies of adsorption from data on reversed-phase chromatography. The authors were able to predict the elution order of nine derivatized peptides that covered a wide range of structures. In 2003, we introduced an artificial neural networks (ANNs) method for predicting peptide elution times[43] that was originally based on amino acid composition and later extended to include partial peptide sequence information.[44]

We have previously reported an accurate mass and time (AMT) tag proteomics approach that uses accurate mass measurements in conjunction with observed peptide retention time information to more confidently identify peptides. [45−48] Palmblad et al.[49,50] have more recently shown that retention time prediction can be combined with accurate mass measurements to improve proteomics measurements; however, their peptide elution time prediction error was high, possibly due to the limitations of the retention coefficient approach used. In various applications, we have shown that when peptide retention time prediction was combined with peptide/protein identification programs such as SEQUEST, the number of false positive identifications could be decreased and/or the number of confident peptide identifications from LC−MS/MS experiments[51−53] increased. Le Bihan et al.[54] used peptide

(16) Yoshida, T. *J. Chromatogr., A* **1998**, *811*, 61−67.
(17) Yoshida, T.; Okada, T. *J. Chromatogr., A* **1999**, *841*, 19−32.
(18) Hearn, M. T. W.; Aguilar, M. I. *J. Chromatogr.* **1987**, *392*, 33−49.
(19) Petritis, K.; Brussaux, S.; Guenu, S.; Elfakir, C.; Dreux, M. *J. Chromatogr., A* **2002**, *957*, 173−185.
(20) Houghten, R. A.; Ostresh, J. M. *BioChromatography* **1987**, *2*, 80−84.
(21) Terabe, S.; Konaka, R.; Inouye, K. *J. Chromatogr.* **1979**, *172*, 163−177.
(22) O'Hare, M. J.; Nice, E. C. *J. Chromatogr.* **1979**, *171*, 209−221.
(23) Wehr, C. T.; Correia, L.; Abbott, S. R. *J. Chromatogr. Sci.* **1982**, *317*, 129−135.
(24) Su, S. J.; Grego, B.; Niven, B.; Hearn, M. T. W. *J. Liq. Chromatogr.* **1981**, *4*, 1745−1753.
(25) Zhou, N. E.; Mant, C. T.; Hodges, R. S. *Pept. Res.* **1990**, *3*, 8−20.
(26) Blondelle, S. E.; Buttner, K.; Houghten, R. A. *J. Chromatogr.* **1992**, *625*, 199−206.
(27) Buttner, K.; Pinilla, C.; Appel, J. R.; Houghten, R. A. *J. Chromatogr.* **1992**, *625*, 191−198.
(28) Sereda, T. J.; Mant, C. T.; Sonnichsen, F. D.; Hodges, R. S. *J. Chromatogr., A* **1994**, *676*, 139−153.
(29) Su, J. Y.; Hodges, R. S.; Kay, C. M. *Biochemistry* **1994**, *33*, 15501−15510.
(30) Rothemund, S.; Krause, E.; Beyermann, M.; Dathe, M.; Engelhardt, H.; Bienert, M. *J. Chromatogr., A* **1995**, *689*.
(31) Sereda, T. J.; Mant, C. T.; Hodges, R. S. *J. Chromatogr., A* **1995**, *695*, 205−221.
(32) Blondelle, S. E.; Ostresh, J. M.; Houghten, R. A.; Perez-Paya, E. *Biophys. J.* **1995**, *68*, 351−359.
(33) Wimley, W. C.; Creamer, T. P.; White, S. H. *Biochemistry* **1996**, *35*, 5109−5124.
(34) Steer, D. L.; Thompson, P. E.; Blondelle, S. E.; Houghten, R. A.; Aguilar, M. I. *J. Pept. Res.* **1998**, *51*, 401−412.
(35) Yu, Y. B.; Wagschal, K. C.; Mant, C. T.; Hodges, R. S. *J. Chromatogr., A* **2000**, *890*, 81−94.
(36) Wieprecht, T.; Rothemund, S.; Bienert, M.; Krause, E. *J. Chromatogr., A* **2001**, *912*, 1−12.
(37) Chen, Y.; Mant, C. T.; Hodges, R. S. *J. Chromatogr., A* **2003**, *1010*, 46−61.
(38) Krokhin, O. V.; Craig, R.; Spicer, V.; Ens, W.; Standing, K. G.; Beavis, R. C.; Wilkins, J. A. *Mol. Cell. Proteomics* **2004**, *3*, 908−919.
(39) Liu, H. X.; Xue, C. X.; Zhang, R. S.; Wao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1979−1986.
(40) Kaliszan, R.; Baczek, T.; Cimochowska, A.; Juszczyk, P.; Wisniewska, K.; Grzonka, Z. *Proteomics* **2005**, *5*, 409−415.
(41) Baczek, T.; Wiczling, P.; Marszall, M.; Heyden, Y. V.; Kaliszan, R. *J. Proteome Res.* **2005**, *4*, 555−563.
(42) Makrodimitris, K.; Fernandez, E. J.; Woolf, T. B.; O'Connell, J. P. *Anal. Chem.* **2005**, *77*, 1243−1252.
(43) Petritis, K.; Kangas, L. J.; Ferguson, P. L.; Anderson, G. A.; Pasa-Tolic, L.; Lipton, M. S.; Auberry, K. J.; Strittmatter, E. F.; Shen, Y.; Zhao, R.; Smith, R. D. *Anal. Chem.* **2003**, *75*, 1039−1048.
(44) Petritis, K.; Kangas, L. J.; Strittmatter, E. F.; Xu, Y.; Yan, B.; Camp II, D. G.; Lipton, M. S.; Smith, R. D. 52nd ASMS conference on Mass Spectrometry and Allied Topics, Nashville, TN, 2004; poster.
(45) Conrads, T. P.; Anderson, G. A.; Veenstra, T. D.; Pasa-Tolic, L.; Smith, R. D. *Anal. Chem.* **2000**, *72*, 3349−3354.
(46) 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.
(47) Lipton, M. S.; Pasa-Tolic, 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.; Markillie, L. M.; Moore, R.; Romine, M. F.; Shen, Y.; Strittmatter, E.; Tolic, N.; Udseth, H. R.; Venkateswaran, A.; Wong, K. K.; Zhao, R.; Smith, R. D. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11049−011054.
(48) Strittmatter, E. F.; Ferguson, P. L.; Tang, K.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 980−991.
(49) Palmblad, M.; Ramstrom, M.; Markides, K. E.; P., H.; Bergquist, J. *Anal. Chem.* **2002**, *74*, 5826−5830.
(50) Palmblad, M.; Ramstrom, M.; Bailey, G. B.; McCutchen-Maloney, S. L.; Bergquist, J.; Zeller, L. C. *J. Chromatogr., B* **2004**, *803*, 131−135.
(51) Strittmatter, E. F.; Kangas, L. J.; Petritis, K.; Mottaz, H. M.; Anderson, G. A.; Shen, Y.; Jacobs, J. M.; Camp, D. G., 2nd; Smith, R. D. *J. Proteome Res.* **2004**, *3*, 760−769.
(52) Varnum, S. M.; Covington, C. C.; Woodbury, R. L.; Petritis, K.; Kangas, L. J.; Abdullah, M. S.; Pounds, J. G.; Smith, R. D.; Zangar, R. C. *Breast Cancer Res. Treat.* **2003**, *80*, 87−97.
(53) Qian, W. J.; Liu, T.; Monroe, M. E.; Strittmatter, E. F.; Jacobs, J. M.; Kangas, L. J.; Petritis, K.; Camp, D. G., 2nd; Smith, R. D. *J. Proteome Res.* **2005**, *4*, 53−62.

elution time prediction parameters to build an empirical model for predicting peptides that are likely to be observable by LC−MS/MS; the model was used for targeted mass spectrometric identification of low-abundance proteins in complex protein samples. Kawakami et al.[55] developed a program that validates peptide assignments based solely on the correlation between the measured and predicted LC elution time of each peptide. In a recent publication, Norbeck et al.[56] demonstrated how accurate mass and normalized elution time (NET) information improved peptide identifications in the study of proteomes of high complexity. Such improvements can significantly extend the protein coverage of highly confident peptide identifications. Similarly, Cargile et al.[57−59] demonstrated confident peptide identifications could be further enhanced by the application of information from isoelectric focusing fractionation as a first dimension in shotgun proteomics. The good correlations observed between predicted and experimental peptide pI values allowed pI information to be used as an additional filtering step to increase the confidence of peptide/protein identifications.

In the model development reported herein, we have explored various approaches for increasing peptide elution time prediction accuracy in RPLC. In addition to more complex ANN architectures, we examined several peptide physicochemical (peptide length, hydrophobicity, etc.) and sequence-dependent parameters (peptide sequence, amphipathicity, nearest neighbor, etc.) that have been shown to affect the peptide retention time in LC. The predictive capability of the model was evaluated by comparing it with several other previously described peptide retention time prediction models. The result shown here has been a significant improvement in predictive capability.

EXPERIMENTAL SECTION

Sample Preparation of Bacterial Tryptic Peptides. Peptide identifications from a number of different bacterial organisms and from an array of studies were used to train and test the ANN. Table 2 lists the bacteria and cites published studies providing the detailed sample preparation for each organism.[60−70] In general, bacterial cells were cultured in tryptone, glucose, and yeast extract medium to an approximate optical density of 600 nm and harvested by centrifugation at 10000*g* at 4 °C. Prior to lysis, cells were resuspended and washed 3 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, and incubated on ice for 5 min between each cycle. The supernatant containing soluble cytosolic proteins was recovered following centrifugation at 15000*g* for 15 min to remove cell debris. Proteins were denatured and reduced in 50 mM Tris buffer (pH 8.2), 8 M urea, 10 mM tributyl phosphine for 1 h at 37 °C. The protein sample was diluted 10 times using 20 mM Tris buffer (pH 8.2) and then digested overnight at 37°C using sequencing grade, modified porcine trypsin (Promega, Madison, WI) at a trypsin/protein ratio of 1:50. The digests were purified using SPE C18 columns (Supelco, Bellefonte, PA) according to the manufacturer's instructions and dried under vacuum.

Table 1. Filtering Criteria (X_{Corr} Thresholds) Used To Select Development Data[a]

	LCQ		LTQ	
charge state and MW	partially tryptic	fully tryptic	partially tryptic	fully tryptic
CS +1, MW <1000	NO	1.6	NO	1.7
CS +1, MW >1000	2.8	2.2	2.9	2.3
CS +2, MW any	3	2.2	4.3	2.4
CS +3, MW any	3.7	2.9	4.7	3.2

[a] The criteria are different depending on the ion trap instrument, the charge state of the peptides, and the peptide molecular weight (in the case of singly charged peptides no partially tryptic peptides with MW <1000 were used).

Preparation of Yeast (*Saccharomycess cerevisiae*) Protein Digests. *S. cerevisiae* (ATCC 26108, Lot 137504) was grown in a batch shaker flask at 37 °C on yeast nitrogen base without amino acids. Medium was prepared with the addition of 5 g/L glucose and 5 g/L fructose. Cells were harvested at mid-logarithmic and stationary phases by centrifugation at 4000 rpm for 10 min. Cells were combined in a ratio of 1:3 stationary-phase cells to mid-logarithm cells. Half of the cells were resuspended in 4 pellet volumes of a denaturation solution (7 M urea, 2 M

(54) Le Bihan, T.; Robinson, M. D.; Stewart, I. I.; Figeys, D. *J. Proteome Res.* **2004**, *3*, 1138−1148.
(55) Kawakami, T.; Tateishi, K.; Yamano, Y.; Ishikawa, T.; Kuroki, K.; Nishimura, T. *Proteomics* **2005**, *5*, 856−864.
(56) Norbeck, A. D.; Monroe, M. E.; Adkins, J. N.; Anderson, K. K.; Daly, D. S.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 1239−1249.
(57) Cargile, B. J.; Bundy, J. L.; Freeman, T. W.; Stephenson, J. L., Jr. *J. Proteome Res.* **2004**, *3*, 112−119.
(58) Cargile, B. J.; Stephenson, J. L., Jr. *Anal. Chem.* **2004**, *76*, 267−275.
(59) Cargile, B. J.; Talley, D. L.; Stephenson, J. L., Jr. *Electrophoresis* **2004**, *25*, 936−945.
(60) Jacobs, J. M.; Yang, X. H.; Luft, B. J.; Dunn, J. J.; Camp, D. G.; Smith, R. D. *Proteomics* **2005**, *5*, 1446−1453.
(61) Varnum, S. M.; Streblow, D. N.; Monroe, M. E.; Smith, P.; Auberry, K. J.; Pasa-Tolic, L.; Wang, D.; Camp, D. G.; Rodland, K.; Wiley: S.; Britt, W.; Shenk, T.; Smith, R. D.; Nelson, J. A. *J. Virol.* **2004**, *78*, 13395−13395.
(62) Lipton, M. S.; Pasa-Tolic, 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.; Markillie, L. M.; Moore, R. J.; Romine, M. F.; Shen, Y.; Stritmatter, E.; Tolic, N.; Udseth, H. R.; Venkateswaran, A.; Wong, K. K.; Zhao, R.; Smith, R. D. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11049−11054.
(63) Ding, Y.-H. R.; Hixson, K. K.; Giometti, C. S.; Stanley, A.; Esteve-Núñez, A.; Khare, T.; Tollaksen, S. L.; Zhu, W.; Adkins, J. N.; Lipton, M. S.; Smith, R. D.; Mester, T.; Lovely, D. R. *BBA Proteins Proteomics*, in press.
(64) Liu, T.; Qian, W.-J.; Chen, W.-N. U.; Jacobs, J. M.; Moore, R. J.; Anderson, D. J.; Gritsenko, M. A.; Monroe, M. E.; Thrall, B. D.; David G. Camp, I.; Smith, R. D. *Proteomics* **2005**, *5*, 1263−1273.
(65) Jacobs, J. M.; Mottaz, H. M.; Yu, L. R.; Anderson, D. J.; Moore, R. J.; Chen, W. U.; Auberry, K. J.; Strittmatter, E. F.; Monroe, M. E.; Thrall, B. D.; Camp, D. G.; Smith, R. D. *J. Proteome Res.* **2004**, *3*, 68−75.
(66) Purvine, S.; Picone, A. F.; Kolker, E. *Omics−J. Integr. Biol.* **2004**, *8*, 79−92.
(67) Callister, S. J.; Nicora, C. D.; Zeng, X.; Roh, J.; Dominguez, M.; Tavano, C.; Monroe, M. E.; Kaplan, S.; Donohue, T.; Smith, R. D.; Lipton, M. S. *J. Microbiol. Methods*, in press.
(68) Prokisch, H.; Scharfe, C.; Camp, D. G.; Xiao, W. Z.; David, L.; Andreoli, C.; Monroe, M. E.; Moore, R. J.; Gritsenko, M. A.; Kozany, C.; Hixson, K. K.; Mottaz, H. M.; Zischka, H.; Ueffing, M.; Herman, Z. S.; Davis, R. W.; Meitinger, T.; Oefner, P. J.; Smith, R. D.; Steinmetz, L. M. *Plos Biol.* **2004**, *2*, 795−804.
(69) Adkins, J. N.; Mottaz, H. M.; Norbeck, A. D.; Rue, J.; Clauss, T.; Purvine, S.; Heffron, F.; Smith, R. D. *Mol. Cell. Proteomics*, in press.
(70) Kolker, E.; Picone, A. F.; Galperin, M. Y.; Romine, M. F.; Higdon, R.; Makarova, K. S.; Kolker, N.; Anderson, G. A.; Qiu, X. Y.; Auberry, K. J.; Babnigg, G.; Beliaev, A. S.; Edlefsen, P.; Elias, D. A.; Gorby, Y. A.; Holzman, T.; Klappenbach, J. A.; Konstantinidis, K. T.; Land, M. L.; Lipton, M. S.; McCue, L. A.; Monroe, M.; Pasa-Tolic, L.; Pinchuk, G.; Purvine, S.; Serres, M. H.; Tsapin, S.; Zakrajsek, B. A.; Zhou, J. H.; Larimer, F. W.; Lawrence, C. E.; Riley, M.; Collart, F. R.; Yates, J. R.; Smith, R. D.; Giometti, C. S.; Nealson, K. H.; Fredrickson, J. K.; Tiedje, J. M. *Proc. Natl. Acad. Sci.U.S.A.* **2005**, *102*, 2099−2104.

Table 2. Organisms from Which the Peptides Were Identified, Number of LC−MS/MS Analyses for Each Organism, Number of Redundant Peptides Identified from Each Organism before Any Filtering, and Number of Different Peptides Used from Each Organism after Filtering with the Criteria of Table 1[a]

organism	no. of LC−MS/MS runs	total peptide IDs	unique filtered peptides	ref
Arabidopsis thaliana	36	8 567	2 049	unpublished
Borrelia burgdorferi	186	145 067	6 945	Jacobs et al.[60]
bovine serum albumin	45	5 853	32	na
cytomegalovirus	125	88 166	3 342	Varnum et al.[61]
Deinococcus radiodurans	1063	491 437	21 912	Lipton et al.[62]
Desulfovibrio desulfuricans	426	624 901	28 826	unpublished
Desulfovibrio vulgaris	66	49 332	650	in preparation
Escherchia coli	16	7 247	126	unpublished
Geobacter metallireducens	116	400 292	21 509	unpublished
Geobacter sulfurreducens	791	909 730	26 446	Ding et al.[63]
Homo sapiens	1254	523 142	31 505	Liu et al.,[64] Jacobs et al.[65]
Mus musculus	697	570 471	34 579	in preparation
Plasmodium falciparum	21	73 421	7 059	unpublished
protein standard mixture[b]	1067	1 183 116	1 154	Purvine et al.[66]
Rhodobacter sphaeroides	1062	432 450	22 766	Callister et al.[67]
Rhodopseudomonas palustris	131	15 750	4 433	unpublished
Saccharomyces cerevisia	606	286 528	12 035	Prokisch et al.[68]
Salmonella typhi	418	1 353 968	27 411	in preparation
Salmonella typhimurium	492	1 692 917	32 920	Adkins et al.[69]
Shewanella oneidensis	2348	3 040 760	33 480	Kolker et al.[70]
synechocystis	343	274 200	15 185	in preparation
Vaccinia virus	13	27 298	1 546	in preparation
Yersina pestis	737	221 196	10 052	Hixson et al.[71]
total	12 059	12 425 809	345 962	

[a] The references are meant to be representative of the samples and sample preparation methods for each organism; it should be noted that the peptide identifications for each organism may contain some quantity of samples and sample preparations that are currently unpublished. [b] The protein standard mixture contains the same peptides and proteins as described by Purvine et al. 2004.[66]

thiourea, in 50 mM ammonium bicarbonate buffer, pH 7.8). Lysis was achieved by bead beating the cell mixture with 0.1-mm zirconia/silica beads in a minibead beater (Biospec, Bartlesville, OK) for 90 s at 4500 rpm. Lysate was collected and placed immediately on ice to inhibit proteolysis. The other half of the cells were subjected to bead beating with a denaturation solution, in which thiourea was absent.

The lysates were reduced by adding neutralized Tris-2-carboxyethylphosphine (Pierce, Rockford IL) to a final concentration of 5 mM and incubated for 30 min at 60 °C. The lysates were then diluted 10-fold with 50 mM ammonium bicarbonate (pH 7.8), and 1 M calcium chloride was added to a final concentration of 1 mM. Proteolysis was achieved by adding sequencing grade modified trypsin (Promega) in an approximate protease to lysate protein ratio of 1:50. The samples were digested for 5 h at 37°C. The lysate that contained no thiourea was alkylated by adding 195 mM iodoacetamide to a final concentration of 10 mM and incubated at room temperature for 30 min. Finally, the lysates were combined, and the peptides were desalted using Supelco (St. Louis, MO) Supelclean C-18 tubes with a Supelco vacuum manifold.

Preparation of Mouse Brain Tissue and Voxel Samples. Brain tissue samples from C57BL/6J male mice were prepared as previously described.[72] The samples were lysed in 80 μL of 5 mM PBS with 80 μL of TFE with intermittent sonication in an ice−water bath. The lysate was reduced with 5 mM TBP and digested by trypsin overnight, and the digests were lyophilized immediately after digestion without further cleanup. Peptide samples were redissolved in 100 μL of 50 mM NH_4HCO_3, and the peptide concentrations were measured by using the BCA protein assay.

Preparation of Human Mammalian Epithelial Cell Protein Digests. Samples were prepared as described previously.[65]

Nearest-Neighbor Effect. The simplest and most direct way of incorporating the nearest-neighbor effect of the 21 amino acids is to construct either a 21 × 21 or 21 × 21 × 21-dimensional array that includes all 441 or 9261 possible combinations (i.e., AA, AC, AD, ... or AAA, AAC, AAD, ...), respectively. The bipeptides/tripeptides in a given peptide are either counted or structured in the ANN in the same way as they appear in the peptide sequence. Alternatively, it is possible to construct the nearest-neighbor list based on an amino acid property. The 21 amino acids can be divided on the basis of their side-chain properties into five groups: (1) nonpolar aliphatic (AGILPV), (2) polar uncharged (CMNQST and C alkylated), (3) aromatic (FWY), (4) positively charged (HKR), and (5) negatively charged (DE). We used this alternative approach to obtain a reduced 5 × 5 dimensional nearest-neighbor array, which is optimal when the number of training peptides is not large enough.

Quasi-Sequence-Order Approach. Due to the extremely large number of possible amino acid residue sequences, it is

(71) Hixson, K. K.; Adkins, J. N.; Gonzales, A.; Moore, R. J.; Smith, R. D.; McCutchen-Maloney, S. L.; Lipton, M. S. *J. Proteome Res.*, in press.

(72) Brown, V. M.; Ossadtchi, A.; Khan, A. H.; Yee, S.; Lacan, G.; Melega, W. P.; Cherry, S. R.; Leahy, R. M.; Smith, D. J. *Genome Res.* **2002**, *12*, 868−884.

difficult to directly incorporate the amino acid sequence order effectively into a statistical prediction algorithm. As a result, we used the "quasi-sequence-order" approach, first introduced by Chou[73,74] to predict protein subcellular locations and attributes. The idea is to assume that the sequence order effect of L-amino acids with the form $a_1a_2a_3a_4a_5\cdots a_L$, can be approximately reflected through the following set of sequence-order-coupling factors:

$$\tau_1 = \frac{1}{L-1}\sum_{i=1}^{L-1} J_{i,i+1}$$

$$\tau_2 = \frac{1}{L-2}\sum_{i=1}^{L-2} J_{i,i+2}$$

$$\tau_3 = \frac{1}{L-3}\sum_{i=1}^{L-3} J_{i,i+3}$$

$$\vdots$$

$$\tau_\lambda = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} J_{i,i+\lambda}, (\lambda < L) \qquad (1)$$

where τ_1 denotes the first-rank sequence-order coupling factor that reflects the sequence-order correlation among all the most contiguous residues along a peptide sequence, τ_2 is the second-rank sequence-order-coupling factor that reflects the sequence-order correlation between all the second most contiguous residues, and so forth; when $\lambda \geq L$, we assign$\tau_\lambda = 0$. The correlation function is described by

$$J_{i,j} = D^2(a_i,a_j) \qquad (2)$$

where $D(a_i,a_j)$ is the physicochemical evolution distance from amino acid a_i to amino acid a_j that was derived on the basis of the residue properties hydrophobicity, hydrophilicity, polarity, and side-chain volume (see Table 1 of Schneider and Wrede[75]).

Secondary Structure Contributions. To incorporate conformational information that can influence chromatographic behavior, we introduced the predicted secondary structural contents (SSC) for each peptide. The SSC attempts to represent the percentage of a peptide's residues that reside in a secondary structural state, e.g., α-helix, β-sheet, or coil. In this study, two different approaches were used to calculate the SSC. In the first approach, the SSC was predicted from the amino acid composition using the shared program SSCP.[76] In the second approach, the SSC was converted from the secondary structure predicted by SSP, which makes use of profiles generated by the PSI-BLAST program. and the PSIPRED secondary structure prediction method of Jones.[77] Generally, peptides with only sufficient lengths have secondary structures, so the SSP was employed for peptides with at least 15 amino acid residues. For those peptides with residues of <15, we arbitrarily treated them as coil.

Hydrophobic Moment. A known phenomenon that causes retention time shifts for isomer peptides is the amphipathicity of the peptides. The amphiphilic helices are those in which one surface of each helix projects mainly hydrophilic side chains, while the opposite surface projects mainly hydrophobic side chains. To quantify the amphiphilicity of a helix, we applied the hydrophobic moment proposed by Eisenberg et al.[78−80] The mean hydrophobic moment can be calculated for an amino acid sequence of N residues and their associated hydrophobicities H_n with the following equation:

$$\langle\mu_H\rangle = \{[\sum_{n=1}^{N} H_n \sin(2n\pi/3.6)]^2 + [\sum_{n=1}^{N} H_n \cos(2n\pi/3.6)]^2\}^{1/2} \qquad (3)$$

A large value for $\langle\mu_H\rangle$ equates to a large peptide amphipathicity.

Capillary LC Coupled with ESI-MS. HPLC-grade water and acetonitrile were purchased from Aldrich (Milwaukee, WI). Fused-silica capillary columns (30−85 cm, 50−50 μm i.d. × 180−360 μm o.d., Polymicro Technologies, Phoenix, AZ) packed with 3.5-μm C18 Jupiter300 particles (Phenomenex, Torrance, CA) were manufactured in-house as described previously.[81] Capillary RPLC was performed using an ISCO LC system (model 100DM, ISCO, Lincoln, NE), and the mobile phases for the gradient elution consisted of (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 were delivered at 5000−10000 psi, using two ISCO pumps to a stainless steel mixer (~2.5 mL), where they were mixed using a magnetic stirrer. The flow was split prior to entering the separation capillary to generate a nonlinear (exponential) gradient[82] and an analysis separation time of ~100 min. Fused-silica capillary flow splitters (various lengths) were used to control the gradient speed. Capillary RPLC was coupled on-line with MS through an ESI interface (a stainless steel union was used to connect the ESI emitter and the capillary separation column).[82]

The peptide database was generated from analyses performed previously using several mass spectrometers, including 3.5, 7, 9, and 11.4 T capillary LC-FTICR instruments (described elsewhere in detail[83] and in references therein), an LTQ-FT (ThermoFinnigan, San Jose, CA), and LCQ Duo, LCQ Deca, LCQ XP, and LTQ (ThermoFinnigan) ion trap mass spectrometers. The ANN software NeuroWindows Version 4.5 (Ward Systems Group) utilized a standard back-propagation algorithm on a Pentium 3.0-GHz personal computer.

RESULTS AND DISCUSSION

In this study, parameters that have been shown to affect the peptide retention time in LC were examined, to investigate their

(73) Chou, K. C. *Biochem. Biophys. Res. Commun.* **2000**, *278*, 477−483.
(74) Chou, K. C. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 246−255.
(75) Schneider, G.; Wrede, P. *Biophys. J* **1994**, *66*, 335−344.
(76) Eisenhaber, F.; Imperiale, F.; Argos, P.; Frommel, C. *Proteins: Struct., Funct., Genet.* **1996**, *25*, 157−168.
(77) Jones, D. T. *J. Mol. Biol.* **1999**, *292*, 195−202.
(78) Eisenberg, D.; Weiss, R. M.; Terwillinger, T. C. *Nature* **1982**, *299*, 371−374.
(79) Eisenberg, D. *Annu. Rev. Biochem.* **1984**, *53*, 595−623.
(80) Eisenberg, D.; Weiss, R. M.; Terwillinger, T. C. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 140−144.
(81) Shen, Y.; Zhao, R.; Belov, M. E.; Conrads, T. P.; Anderson, G. A.; Tang, K.; Pasa-Tolic, L.; Veenstra, T. D.; Lipton, M. S.; Smith, R. D. *Anal. Chem.* **2001**, *73*, 1766−1775.
(82) Shen, Y.; Tolic, N.; Zhao, R.; Pasa-Tolic, L.; Li, L.; Berger, S. J.; Harkewicz, R.; Anderson, G. A.; Belov, M. E.; Smith, R. D. *Anal. Chem.* **2001**, *73*, 3011−3021.
(83) Harkewicz, R.; Belov, M. E.; Anderson, G. A.; Pasa-Tolic, L.; Masselon, C. D.; Prior, D. C.; Udseth, H. R.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 144−154.

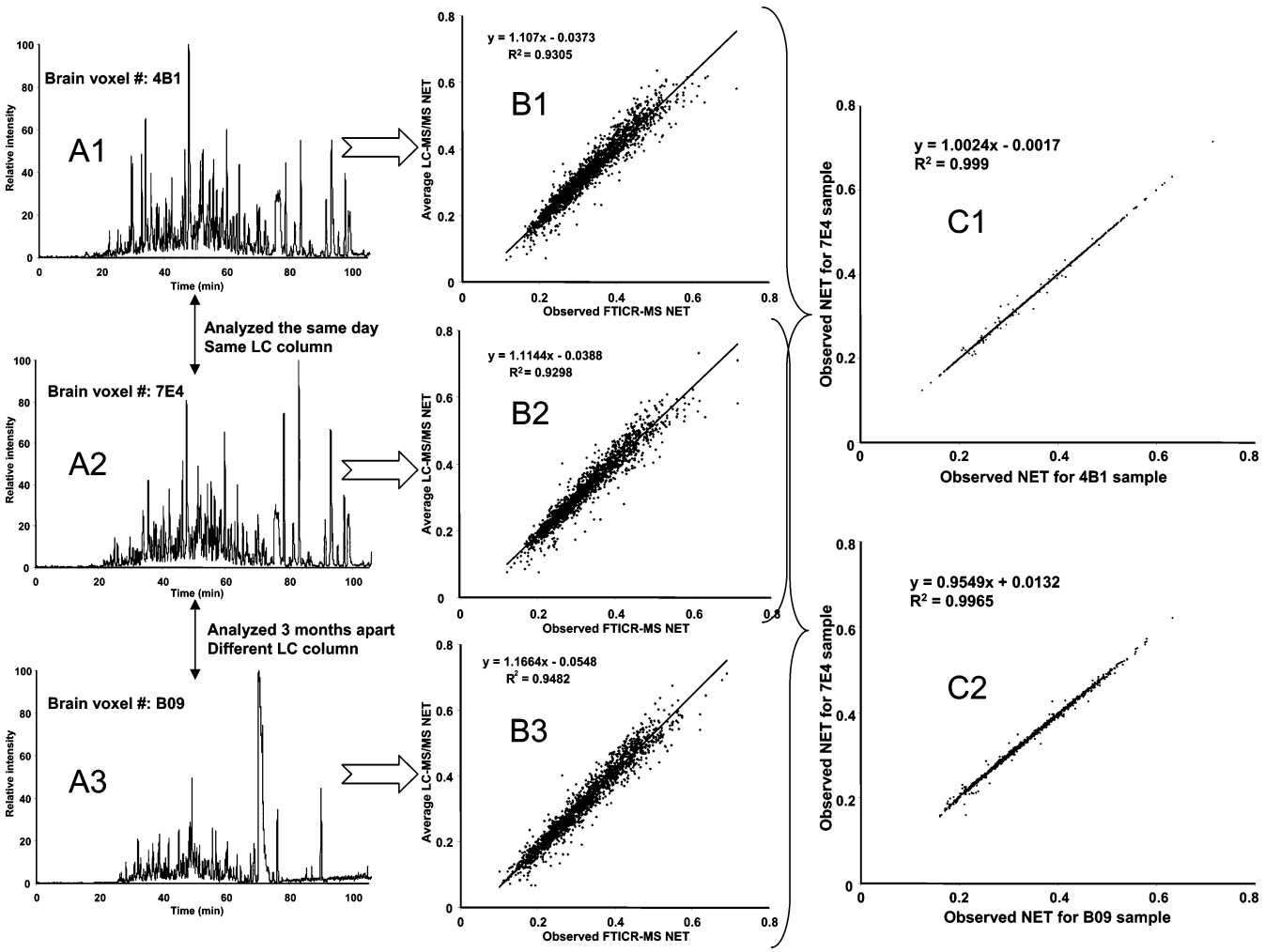


Figure 1. Schematic representation of the present study normalization method. (A1−3) Figure depicts the base peak chromatograms of given mouse brain voxels from different brain regions of the same mouse. The voxels 4B1 and 7E4 were analyzed by LC−FTICR MS on the same day and with the same chromatographic column. The voxels B09 was analyzed 3 months later using different chromatographic column (but having the same dimensions and chromatographic packing). (B1−3) Observed accurate mass and time are regressed against computed masses and average observed NET values from LC−MS/MS using an iterative process. The regression residual converges when the observed accurate mass and time match their theoretical/predicted ones. The slope and intercept of the trendline are used for the linear (regression) based mapping of observed elution time to observed NET (C1,2) These plots show the correlation of observed NET values for the peptides in common between different LC−FTICR MS analyses.

incorporation in approaches for improvement in the predictive capability of the model. Some of these values (i.e., hydrophobic moment, secondary structure, etc.) are calculated/predicted values, while others (i.e., length, sequence, etc.) are known values—as long as the peptide identification is correct—that have been encoded in the model by using more complex artificial neural networks. In the case of full encoding of the peptide sequence, a large number of peptide identifications were necessary for the training set. As a result, for reasons described later, new filtering criteria for selection of the most confidently identified peptides as well as improved peptide LC elution time normalization procedures were needed.

Normalization of Peptide LC Elution Times. The ANN and training algorithms employed for the present model development effort were described in our previous relevant work.[43] Briefly, peptides identified from the radiation-resistant organism *Deinococcus radiodurans* (~7000 peptides) were used to train the ANN, and peptides identified from the metal-reducing organism *Shewanella oneidensis* (~5200) were used to test it. A genetic algorithm developed to normalize the peptide elution times into a range (from 0 to 1) and correlate data sets enabled accurate comparison of numerous LC−MS data sets and improved the peptide elution time reproducibility to ~1%. This algorithm was based on a linear regression of a set of six peptides identified frequently in both organisms of the study[43] and normalized peptide elution times coming from both the same and different organisms.

While generating excellent results, this normalization approach became time prohibitive as the number of peptides used increased significantly. To train/test the present model, we employed confidently identified peptides from 22 different organisms, as well as a mixture of standard proteins/peptides,[66] that provided a set of ~12 million peptide identifications. The time needed to normalize this set of peptide identifications using the genetic algorithm (described above) would be on the order of several weeks due to the many generations (iterations) required to align all analyses through multiple regressions.[84,85] Consequently, we revised the process to normalize each LC−MS/MS analysis independently by regressing all the observed peptide elution times

in a given analysis against the predicted NET for the same peptides. Although we initially constructed the scale of predicted NET values by using the previously trained genetic algorithm, once the algorithm had been trained, we were able to use the NET values predicted by this algorithm for LC−MS/MS alignment. The alignment of each LC−MS/MS data set against the list of predicted NET values provided the means to convert the observed elution time for each peptide to an observed NET value on the basis of the relationship $NET_{observed}$ = (slope × elution $time_{observed}$) + intercept.

A step in the accurate mass and time (AMT) tag proteomics approach developed in our laboratory involves using peptide observations from multiple (and often extensive sets of) LC−MS/MS analyses of appropriately related samples[45,46] to create a reference database of accurate mass and observed LC NET values for each identified peptide. These AMT tags are used to identify peptides in subsequent high-throughput LC−MS analyses of the same organism. For peptides observed in several LC−MS/MS analyses, the observed NET values are averaged, which provides statistics on the distribution of NET values for each peptide. In analyses by LC−MS, e.g., LC-Fourier transform ion cyclotron resonance (FTICR), data consist of a list of observed peptide "features," wherein each feature consists of a monoisotopic mass (after collapse of the isotopic distribution for the peptide and subtraction of the proton(s) mass) and an observed elution time. To derive NET values for the detected features, we used an iterative process to regress the observed accurate mass and elution time against the computed masses and averaged observed NETs in the reference database. The regression residual converges when the observed accurate mass and elution time match their theoretical/predicted ones. The slope and intercept of the trend line are used for linear (regression)-based mapping of observed elution time to observed NET. Figure 1 shows some representative "real-world" data of the present normalization method applied to LC-FTICR experiments of mouse brain voxels. Figure 1A depicts the base peak chromatograms of given mouse brain voxels from different spatial brain sections of the same mouse. The voxels 4B1 and 7E4 were analyzed by LC−FTICR-MS on the same day while the voxel B09 was analyzed 3 months later in a different chromatographic column of the same dimensions. Figure 1B shows the correlations obtained when observed accurate mass and time are regressed against computed masses and average observed NET values from LC−MS/MS using an iterative process. Finally, Figure 1C of this figure shows that the NETs of peptides in common among LC-FTICR analyses are highly correlated (R^2 > 0.99), even for experiments that were performed three months apart. By normalizing the elution time of all peptides, we optimize the overall alignment of both LC−MS and LC−MS/MS data sets, an important step for more effective peptide identification[45,46] and quantitation using the AMT tag approach.[86]

Peptide Identification Data for the Training and Testing of the Artificial Neural Network Model. Our earlier work[43] was limited by both the uncertain levels of confidence associated with peptide identifications and the relatively small number of different peptides. However, by using different organisms to train and test the ANN model, we demonstrated that the earlier model was unbiased toward the peptides of a specific organism, and therefore, peptides from any organism could be used to populate the training/testing database.

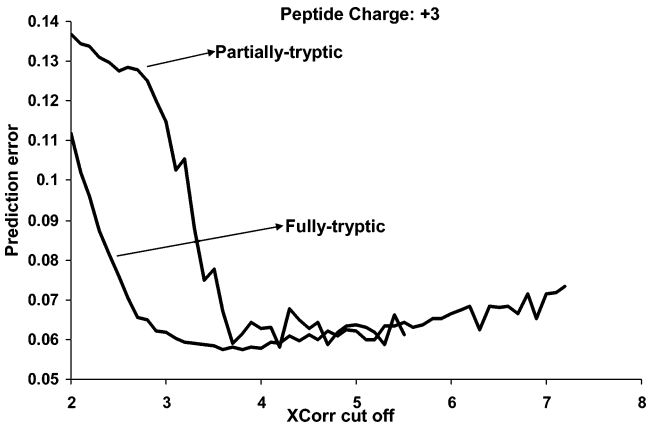


Figure 2. Peptide retention time prediction error distribution vs peptide X_{corr} values for partial and fully tryptic triply charged peptides run on a ThermoFinnigan LCQ ion trap. The filtering criteria given in Table 1 were generated based upon plots.

One of the main objectives of this study was to incorporate peptide sequence information into an ANN architecture, and it was evident from the start that a large training set would be required. Based on our experience, the best way to obtain a large number of new peptide structures was by analyzing different organisms. We eliminated the filtering requirement of ≥3 identifications per peptide so that peptides from "new" organisms (that had not been analyzed multiple times) were included in the data set, and we also changed our filtering criteria. A peptide database of ~12 million redundant peptides identified by SEQUEST from LC−MS/MS analyses of tryptically digested proteomes for an array of organisms was assembled and used to calculate a new set of criteria that provided the best correlation between observed and predicted peptide NETs. A minimum of five amino acid residues was required for each peptide identification, and the data were filtered to include only those peptides with $X_{corr} \geq 1.5$ for a peptide mass of <1000 Da and $X_{corr} \geq 2.0$ for a peptide mass of ≥1000 Da. The filtered peptides were separated into categories according to their charge (1+, 2+, 3+), tryptic state (fully and partial tryptic), and ion trap MS analyzer (LCQ or LTQ). In the case of singly charged peptides, peptides were further categorized on the basis of mass, i.e., MW <1000 Da and MW ≥1000.

For each category, we calculated the elution time prediction error versus the different peptide X_{corr} values by using one of our previously developed peptide elution time predictors.[44] Figure 2 illustrates these calculations for triply charged peptides, analyzed by LC−ion trap (LCQ) MS for fully and partial tryptic peptides. The X_{corr} thresholds were set to values that provided good correlations between observed and predicted NETs. Table 1 summarizes the X_{corr} threshold for each peptide category. Note that higher X_{corr} threshold values were needed for the LTQ-based analyses than for the LCQ. This finding might be attributed to

(84) Holland, J. H. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, MI, 1975.
(85) Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: Reading. MS, 1989.
(86) Qian, W. J.; Jacobs, J. M.; Camp, D. G., II; Monroe, M. E.; Moore, R. J.; Gritsenko, M. A.; Calvano, S. E.; Lowry, S. F.; Xiao, W.; Moldawer, L. L.; Davis, R. W.; Tompkins, R. G.; Smith, R. D. *Proteomics* **2005**, *5*, 572−584.

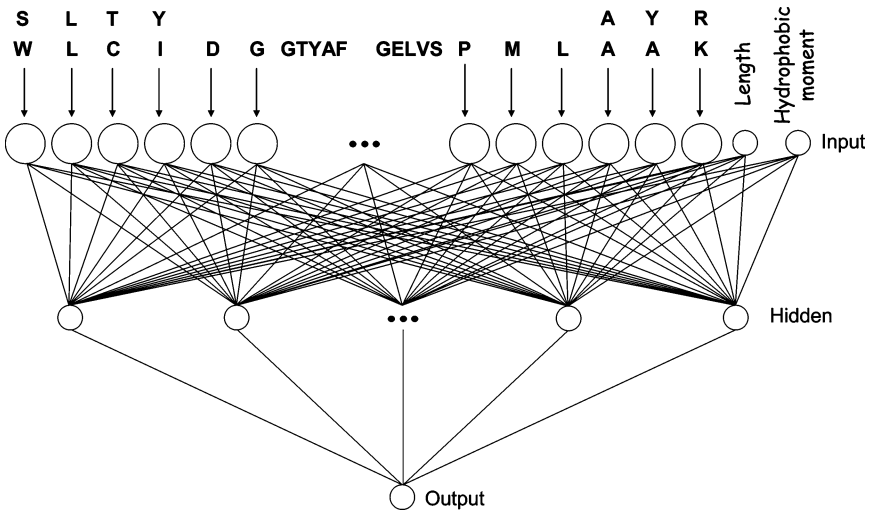


Figure 3. Schematic representation of the artificial neural network architecture used in this study (1052 input nodes, 24 hidden nodes, 1 output node). The large circles represent 21 length vectors while the smaller circles represent single scalar inputs. The small black circles (middle) are used to show continuance.

the better signal-to-noise ratios provided by LTQ mass spectrometers.[87] For singly charged ions, lower X_{corr} threshold values worked better with MW <1000 than with MW ≥1000. This finding may be potentially attributed to the known bias of the SEQUEST algorithm toward peptide mass.[88]

Table 2 provides the organism from which peptides were identified, the number of LC−MS/MS analyses for each organism, the number of unfiltered redundant peptides identified from each organism, and the number of unique peptides identified from each organism after filtering with the criteria used to train/test our model (Table 1). Note that the 12 059 LC−MS analyses generated 345 965 unique filtered peptides for training/testing the present model.

Improvement of the Peptide Elution Time Prediction by Incorporating Peptide Sequence and Conformation Information. Our previous ANN peptide elution time prediction model[43] was based solely on amino acid composition, but had the added advantage over other similar models in that the ANN architecture could better handle nonlinearities. To further improve the peptide elution time prediction, we explored incorporation of several sequence/structural peptide descriptors, including peptide length, sequence, predicted secondary conformation (i.e., helix, sheet, or coil), and hydrophobic moment. In addition to the 20 proteinogenic amino acids, we added alkylated cysteine, since cysteines are reduced and alkylated in most of our mammalian proteomic research.

The first peptide descriptors tested were length and hydrophobic moment. Added to our previous ANN architecture as two additional inputs, these descriptors provided a slight improvement in predictive capability. This improvement is evidenced by looking at the first four rows in Table 3, which show that the correlation between predicted and observed peptide elution time increased from 0.870 to 0.884. Next, we investigated the effect of incorporating peptide sequence into the model by using the quasi-sequence-order approach (see Experimental Section) to describe peptide sequence, and the results were compared with our previous ANN model.[43] This approach did not provide any noticeable improvement over our previous model, so we searched for alternative approaches.

Table 3. Improvement in Peptide Retention Time Prediction with Implementation of Sequence Information, Hydrophobic Moment, and Length of the Peptide in the ANN Model[a]

encoding	hidden	length	hydro moment	train rmse	test rmse	R^2
0/0	4	no	no	0.050 575	0.057 994	0.870 11
0/0	4	no	yes	0.050 504	0.057 678	0.871 35
0/0	4	yes	no	0.048 854	0.055 177	0.879 91
0/0	4	yes	yes	0.048 153	0.054 39	0.883 85
1/1	6	yes	yes	0.044 673	0.052 086	0.892 4
2/2	6	yes	yes	0.040 411	0.045 895	0.916 32
3/3	7	yes	yes	0.038 277	0.042 905	0.926 72
4/4	7	yes	yes	0.036 746	0.040 275	0.935 42
5/5	10	yes	yes	0.035 007	0.037 347	0.944 25
6/6	10	yes	yes	0.034 179	0.036 939	0.945 02
7/7	12	yes	yes	0.033 143	0.035 445	0.949 51
8/8	12	yes	yes	0.032 658	0.034 555	0.951 96
9/9	14	yes	yes	0.031 793	0.034 251	0.953 22
10/10	14	yes	yes	0.031 223	0.033 571	0.954 77
11/11	16	yes	yes	0.031 836	0.033 811	0.953 91
12/12	16	yes	yes	0.031 482	0.033 437	0.955 04
25/25	24	yes	yes	0.026 98	0.028 579	0.966 97

[a] The encoding column refers to the number of amino acid residues defined in the beginning and end of each peptide. The hidden column refers to the number of hidden nodes in the ANN model. rmse, root-mean-square error.

Our prediction models were based on increasingly large quantities of peptide sequence information as the number of data sets grew significantly larger. While increasing the complexity of the ANN model, both in sequence information and in number of hidden nodes, we carefully monitored the process to avoid "overfitting" by using cross-validation during the training process. All results presented here are from ANN models that were trained until both the training and cross-validation errors converged at their lowest values. Thus, early stopping was not necessary, and overfitting was avoided in the final ANN models presented.

(87) Mayya, V.; Rezaul, K.; Cong, Y. S.; Han, D. *Mol. Cell. Proteomics* **2005**, *4*, 214−223.
(88) MacCoss, M. J.; Wu, C. C.; Yates, J. R. *Anal. Chem.* **2002**, *74*, 5593−5599.

Table 4. Number of Times Each Amino Acid Residue Was Found in Different ANN Vector Positions

position	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	c	total
1	36784	427	17302	19797	14445	24812	9551	24733	17751	38513	12302	14690	4581	10247	12433	22767	20633	29556	3071	11116	403	345914
2	35769	664	15483	16818	14642	29751	8804	24804	14799	44663	8459	13213	14401	14169	9028	19028	18301	29004	3049	10468	597	345914
3	33950	701	26954	30939	13129	28076	8595	19865	8603	36461	6425	15016	17603	16400	4612	20127	18801	27187	2999	8891	580	345914
4	32091	737	26887	33189	12629	24592	8460	20413	8042	35520	6269	13746	19203	15104	3946	20460	19885	26414	3095	8645	490	339817
5	30616	775	23724	26915	11853	24655	8000	18149	7826	33509	6485	13829	18128	14236	3753	19164	18248	24322	2942	8847	548	316524
6	27581	825	20687	23214	10768	22709	6837	17110	6999	30278	5681	11430	17410	12143	3366	16410	16297	22661	2397	7557	535	282895
7	23590	679	19189	20819	8671	19210	6046	13806	6461	24261	4547	10351	15331	11079	3043	14631	13905	18518	2295	6173	469	243074
8	19745	585	15508	17026	7337	16238	4782	11577	5494	21029	4091	8595	12806	8996	2541	12258	12161	15739	1835	5319	344	204006
9	16369	579	12709	13690	6094	13540	4239	9751	4473	17520	3324	6705	10856	7215	2214	10209	9606	12975	1368	4409	316	168161
10	13386	502	10656	11239	4808	11192	3216	7825	3736	14474	2796	5693	8734	5729	1760	7921	7917	10501	1175	3539	219	137018
11	11122	411	8331	9123	3877	9104	2497	6284	3088	11299	2314	4481	7075	4699	1335	6456	6544	8394	921	2770	184	110309
12	8943	340	6733	6964	3036	6816	2086	4889	2538	9204	1628	3585	5654	3787	1130	5176	5152	6787	732	2183	134	87497
13	7179	299	5087	5795	2443	5692	1623	3974	1914	7083	1356	2833	4362	2978	882	4157	4000	5105	618	1643	115	69138
14	5592	242	4113	4393	1855	4426	1134	2904	1590	5648	1119	2184	3546	2343	767	3194	3206	4137	462	1356	99	54310
15	4445	196	3078	3314	1467	3232	1031	2471	1183	4324	794	1738	2560	1848	544	2485	2522	3104	334	1043	82	41795
16	3355	126	2253	2472	1024	2541	748	1728	895	3175	667	1333	1945	1336	395	1843	1802	2449	233	839	43	31202
17	2626	98	1588	1824	727	1964	477	1280	589	2375	427	881	1319	1041	321	1425	1286	1686	167	591	27	22719
18	1886	82	1102	1227	464	1316	316	986	409	1688	319	618	994	639	181	992	973	1212	116	355	22	15897
19	1157	43	738	765	324	847	240	620	294	1135	208	386	631	413	102	651	664	884	80	268	13	10463
20	822	21	465	550	220	578	146	384	196	682	104	244	412	268	82	361	404	530	55	140	7	6671
21	483	10	284	293	117	368	76	233	91	404	87	169	237	167	40	236	226	306	24	96	3	3950
22	279	10	141	178	61	193	30	110	51	229	34	62	133	92	19	143	134	151	16	45	2	2113
23	142	3	66	54	15	90	16	44	40	94	23	34	58	35	11	48	64	74	7	19	4	941
24	42	1	25	22	3	32	5	19	5	30	4	6	25	12	4	22	26	28	2	5	0	318
25	12	0	3	6	0	7	0	2	2	3	1	6	3	4	0	5	3	0	0	2	0	59
26	6	0	0	1	2	3	0	2	0	0	0	0	2	0	0	2	0	1	0	0	0	19
27	31	1	6	8	6	13	3	7	5	9	3	3	11	4	0	5	6	13	1	0	1	136
28	95	2	40	47	10	53	4	28	14	61	8	23	39	27	5	34	29	39	1	10	1	570
29	185	7	93	109	39	157	23	76	38	133	19	54	98	59	10	105	78	123	16	33	0	1455
30	383	7	193	190	93	260	58	172	81	313	57	100	203	110	20	191	177	254	16	72	2	2952
31	555	27	334	393	169	491	84	335	143	534	95	209	321	196	56	319	352	414	36	118	10	5191
32	1059	39	647	655	247	655	177	484	206	822	152	354	525	366	105	506	484	645	71	185	11	8395
33	1516	59	885	1027	435	1066	286	756	343	1299	239	459	818	570	168	839	830	1000	89	277	11	12972
34	2143	92	1263	1466	655	1584	380	1154	507	2023	372	723	1248	825	241	1109	1233	1455	138	479	33	19123
35	2872	126	1867	2110	919	2240	561	1446	743	2820	577	1148	1670	1117	356	1574	1605	2049	181	641	43	26665
36	3957	149	2579	3006	1193	3017	738	1946	975	3870	776	1540	2279	1526	478	2107	2187	2681	316	928	58	36306
37	4994	205	3556	3976	1691	3965	1104	2650	1286	4913	1013	1898	3064	1971	588	2762	2880	3592	394	1247	78	47827
38	6336	280	4545	5157	2185	4848	1406	3311	1743	6301	1182	2567	3930	2599	791	3661	3647	4689	512	1507	117	61314
39	8021	283	5947	6794	2645	6310	1818	4368	2302	7971	1599	3114	4860	3340	911	4395	4676	5810	582	2005	119	77870
40	10206	355	7288	8282	3269	7720	2261	5467	2838	10657	2055	4004	6212	4113	1115	5845	5781	7300	889	2551	163	98371
41	12411	445	9153	9972	4228	9698	3135	6847	3340	12996	2391	4978	7912	5131	1445	7486	7462	9628	988	3255	204	123105
42	15150	461	11196	12883	5161	11950	3812	8478	4086	15836	3038	6519	9851	6572	1747	9027	8880	11756	1419	3937	243	152002
43	18355	508	14749	17201	6675	14127	4793	10211	5063	18525	3778	7780	11801	8293	2010	10996	10724	13482	1547	4718	326	185662
44	22765	663	16525	17587	8251	16964	5466	13468	5582	25104	4768	8817	13203	9696	2376	13782	12965	16997	2007	5951	389	223326
45	26918	598	18309	20145	9819	19318	6873	15557	7006	29523	5661	10743	16434	11475	2770	15522	15303	20839	2492	7456	419	263180
46	30892	623	23271	29970	10588	23025	7400	16759	6905	30133	6196	12136	18190	13268	2852	18359	17108	21875	2763	7934	430	300677
47	35339	637	25673	31758	11994	24354	8293	18331	7428	34936	6631	14151	19945	15330	2666	18920	17975	23543	3000	8782	474	330160
48	36919	783	18314	25504	14952	25494	8347	23113	7605	46125	8633	13887	14428	16026	2815	19239	19989	28512	3750	9839	578	344852
49	39340	777	14564	28442	12537	28621	7123	19888	14517	39537	7665	11901	17821	15827	6778	20603	18538	28228	3453	9202	552	345914
50	1626	18	721	1174	870	947	811	492	162551	1579	368	811	333	914	170089	777	388	682	110	615	38	345914

Figure 3 shows a generic diagram of the new ANN architecture, illustrating how peptides were encoded. The new architecture contains 1052 input nodes, 24 hidden nodes, and 1 output node (referred to as model 1052-24-1). Each amino acid residue is coded as a 21-dimensional binary vector that consists of 20 zero values and 1 one value that corresponds to the amino acid residue occupying that position. The length and hydrophobic moment were used as normalized scalar values. In other words, the calculated values of length and hydrophobic moment were normalized to [0−1] ranges and incorporated in the ANN as numerical values.

The amino acid residues were positioned in the ANN, starting from the N and C termini and working toward the center of the vector. Using the 7-residue peptide SLTYAYR as an example, the amino acid residues SLTY are positioned at the first 4 × 21 ANN inputs, and the amino acid residues AYR are positioned at the three last 3 × 21 ANN inputs, leaving the center filled with zero values. Only peptides with 50 amino acid residues fill all of the ANN inputs. Table 4 summarizes the number of times each amino acid residue appeared in different ANN positions. The last column of this table shows the total number of amino acid residues in each position. From this column, readers can extrapolate statistics with regard to the lengths of the peptides in our training/testing database. Table 4 shows that, with the exception of some zero values in the center, there are a significant number of residues per position, even for low-abundant amino acid residues such as Cys and Trp. The same holds true for the 50th position, despite the obvious bias toward Lys and Arg as a result of trypsin, which was used for protein digestion. Furthermore, it should be noted that there are several peptides with Pro appearing as the first amino acid residue, despite the difficulty of trypsin to cleave KP or RP bonds.

The model was tested using 1303 (the highest confident identifications) of the 345 914 peptides identified from more than 90 different LC−MS/MS experiments. The other 344 611 peptides were used for training. Table 3 shows the improvements in peptide retention time prediction due to implementation of increased sequence information, hydrophobic moment, and length of the peptide in the ANN model. Using the same training and testing data sets with our previous ANN model[43] (based solely on amino acid composition), we achieved a correlation coefficient of 0.87. The correlation increased to 0.967 when the full peptide sequence was encoded, and the length and hydrophobic moment were added. Most of the improvement was achieved when at least five amino acid residues were encoded from each side of the peptide (i.e., correlation coefficient increases to 0.944), after which the rate of the improvement slowed. A number of hidden nodes were tested for each residue encoded until an optimal number was determined.

It should be noted that data acquired over ~3 years was used to provide sufficient peptide identifications to fully encode peptides of up to 50 amino acid residues. We first introduced the idea of using peptide sequence information in 2004 showing results for a database of ~98 000 peptides that allowed us to encode 12 amino acids residues at each peptide terminus.[44] We found that encoding amino acid residues that were close to N and C termini provided improved predictions compared to encoding amino acid residues located in the middle of the peptide as shown in Table 5. We performed a sensitivity analysis with the "perturb" method[89] to determine how much each residue position affected the elution time. The method tests how much each input, if perturbed, changes the output of the model, while the other inputs are fixed. We used the testing set of peptides as the fixed inputs to the model. Each input was tested for each peptide, and the sums of these tests were averaged so that the sets of 21 consecutive inputs, representing each residue position, gave us a relative strength of that position. Sensitivity analysis shows that N and C terminus amino acid encoding is more important than the encoding of amino acid residues in the middle of the peptide. This may be because the amino acid residues at the termini of the peptides are more likely to interact with the stationary phase than amino acid residues in the middle of the peptide. Finally, the sensitivity analysis showed that the incorporation of the length and hydrophobic moment in the model is not as important as the incorporation of the peptide sequence.

Table 5. Sensitivity Analysis[89] of Different Variables Used for the Peptide Elution Time Predictor

position	sensitivity analysis
1	0.1375
2	0.1708
3	0.1107
middle average	0.0867
$n-2$	0.1955
$n-1$	0.1993
n	0.2562
length	0.00521
hydrophobic moment	0.00486

To further improve the model, peptide conformational effects were incorporated by adding predicted secondary peptide structural states (α-helix, β-sheet, and coil). However, the addition of these predicted states[76,77] to both the present and earlier versions of the model did not improve the elution time prediction. A possible explanation is that the approaches used to calculate the peptide secondary structural states in this study failed to predict values that adequately simulate the medium that the peptides are dissolved in and their environment during the LC separation (i.e., water/acetonitrile/TFA/acetic acid, acidic pH, hydrophobic stationary phase). As a result, while these values might work for other applications, they failed to improve upon the present model.

We also evaluated the incorporation of information on nearest neighbors into the model. Several different approaches as described in the Experimental Section were investigated. The 5 × 5 dimensional nearest-neighbor list, which divided the amino acids according to their side-chain properties, failed to provide any improvement in our present model, as well as the earlier model. When we incorporated the 21 × 21-dimensional nearest-neighbor list into our earlier model, we observed a significant improvement; i.e., the correlation between observed and predicted elution times increased from 0.87 to 0.91. However, this model was still inferior to the 1052-24-1 model that encodes only the peptide sequence. Fusion of the two models into a single ANN architecture overfits our training set (i.e., insufficient data for training) and would result in poor predictions. The 21 × 21 × 21-dimensional array also

(89) Yao, J.; Teng, N.; Poh, H.-L.; Tan, C. L. *J. Inf. Sci. Eng.* **1998**, *14*, 843−862.

Table 6. Predicted and Observed NET Values of Several Isobaric/Isomeric Peptides

peptide	MW	NET predicted	NET observed	NET abs error
VMAELK	689.3829	0.137 673	0.131 735	0.005 938
MEVLAK	689.3829	0.141 518	0.142 36	0.000 842
NLISLR	714.4435	0.257 648	0.225 998	0.031 65
VILASGR	714.4435	0.131 71	0.142 392	0.010 682
AVGILSR	714.4435	0.184 029	0.179 694	0.004 335
IFEDVK	749.4006	0.165 442	0.164 227	0.001 215
IEFVDK	749.4006	0.182 281	0.173 211	0.009 07
FDVEIK	749.4006	0.202 775	0.191 268	0.011 508
ELMLER	789.4102	0.193 284	0.193 441	0.000 157
ELMELR	789.4102	0.208 153	0.208 77	0.000 617
AMGVDVAK	789.4102	0.127 098	0.118 302	0.008 796
LFQNDPTGR	1046.519	0.133 044	0.132 35	0.000 693
FDGNPQTLR	1046.519	0.152 405	0.154 723	0.002 319
IAFVSTESGNK	1151.587	0.176 421	0.169 248	0.007 173
STIEGFVNASK	1151.587	0.232 176	0.229 142	0.003 034
VLNESTILIFPK	1372.801	0.376 012	0.384 751	0.008 739
VNFLPEIITLSK	1372.801	0.426 017	0.458 545	0.032 528
TIGLGDAAVAEMIR	1415.749	0.361 089	0.390 157	0.029 068
GTGLIAAIEMVADR	1415.749	0.498 76	0.482 163	0.016 598
AGAPQSVDAPLGETVRK	1694.9	0.182 86	0.183 52	0.000 66
KAGAPQSVDAPLGETVR	1694.9	0.191 434	0.193 53	0.002 096
NAALPIFVSTILAPGLNEIR	2108.204	0.563 867	0.575 534	0.011 666
NAALPVFISTILAPGLNEIR	2108.204	0.591 099	0.589 784	0.001 315
IQALEDILDAEHPNWRER	2204.102	0.372 155	0.380 546	0.008 391
ERIQALEDILDAEHPNWR	2204.102	0.401 762	0.408 566	0.006 804
GNYAERVGAGAPVYMAAVLEYLTAEILELAGNAARDNKK	4108.109	0.747 845	0.750 595	0.002 75
KGNYAERVGAGAPVYMAAVLEYLTAEILELAGNAARDNK	4108.109	0.805 271	0.839 432	0.034 16
LKEISYIHAEAYAAGELKHGPLALIDADMPVIVVAPNNELLEK	4654.476	0.508 866	0.511 891	0.003 026
EISYIHAEAYAAGELKHGPLALIDADMPVIVVAPNNELLEKLK	4654.476	0.530 865	0.543 784	0.012 919

overfit our data. Contrary to our expectations, incorporation of the nearest-neighbor effect into our model did not further improve the elution time predictability. We strongly suspect this is because the 1052-24-1 ANN architecture has already implicitly captured the nearest-neighbor information.

In addition to better elution time predictions, the new 1052-24-1 ANN model is also able to more accurately predict isomeric peptide elution times, a capability that no previously published model has accomplished. Previously described predictors[43,90] were able to model separate LC elution times for isobaric peptides but were unable to differentiate the elution times of isomeric peptides. Table 6 shows several examples of accurate predictions among isobaric/isomeric peptides. For example, the isobaric/isomeric peptides NLISKR, VILASGR, and AVGILSR have identical MWs of 714.4435 and are indistinguishable by accurate mass measurements alone. However, because of their different elution times and the ability of the model to accurately predict these elution times, it is now possible to distinguish isobaric/isomeric peptides.

Finally, Table 7 shows the present work has provided a significant improvement in the peptide elution time prediction errors compared with those of our previous ANN model, regardless the length of the peptide. Longer peptides (i.e., 11−40 amino acid residues) show a larger degree of improvement than do very small peptides. This observation is reasonable as the longer the peptide, the more it deviates from the simplistic assumption that elution time depends on the peptide amino acid composition. However, despite the improvements afforded by this study in predicting the elution times of longer peptides, smaller peptides are still predicted with higher precision. Continued population of our database with longer peptides (31−50 residues) will further improve their predicted NETs.

Table 7. Average Mean Square Error (av MSE) of the Peptide Elution Time Prediction in Relation to the Peptide Length

peptide length	peptides with that length	av MSE (Petritis et al. 2003)[43]	av MSE present study
5−10	107	0.000 72	0.000 27
11−20	684	0.002 40	0.000 53
21−30	403	0.005 03	0.001 21
31−40	104	0.005 92	0.001 70
41−50	5	0.004 47	0.002 26

Comparison of Peptide Elution Time Prediction Models. Several peptide elution time prediction models have been described in the past. However, all of them have used different sets of peptides to train their models and most of them did not use a separate set of peptides to test their model (i.e., the model was tested using the training set), making comparison difficult. Furthermore, many of the models used synthetic peptides for training and testing, while others used "real-world" data from mainly proteomic applications, where the potential of false positive identifications filtered through the training and testing set might have negatively affected the reported prediction capability.

For comparison, we decided to train and test several previously reported prediction models[10,13,38,40,43] with the peptide data set used in this study. Among the previously published peptide elution time

(90) Krokhin, O. V.; Craig, R.; Spicer, V.; Ens, W.; Standing, K. G.; Beavis, R.; Wilkins, J. A. 52nd ASMS conference on Mass Spectrometry and Allied Topics, Nashville, TN, 2004; poster.

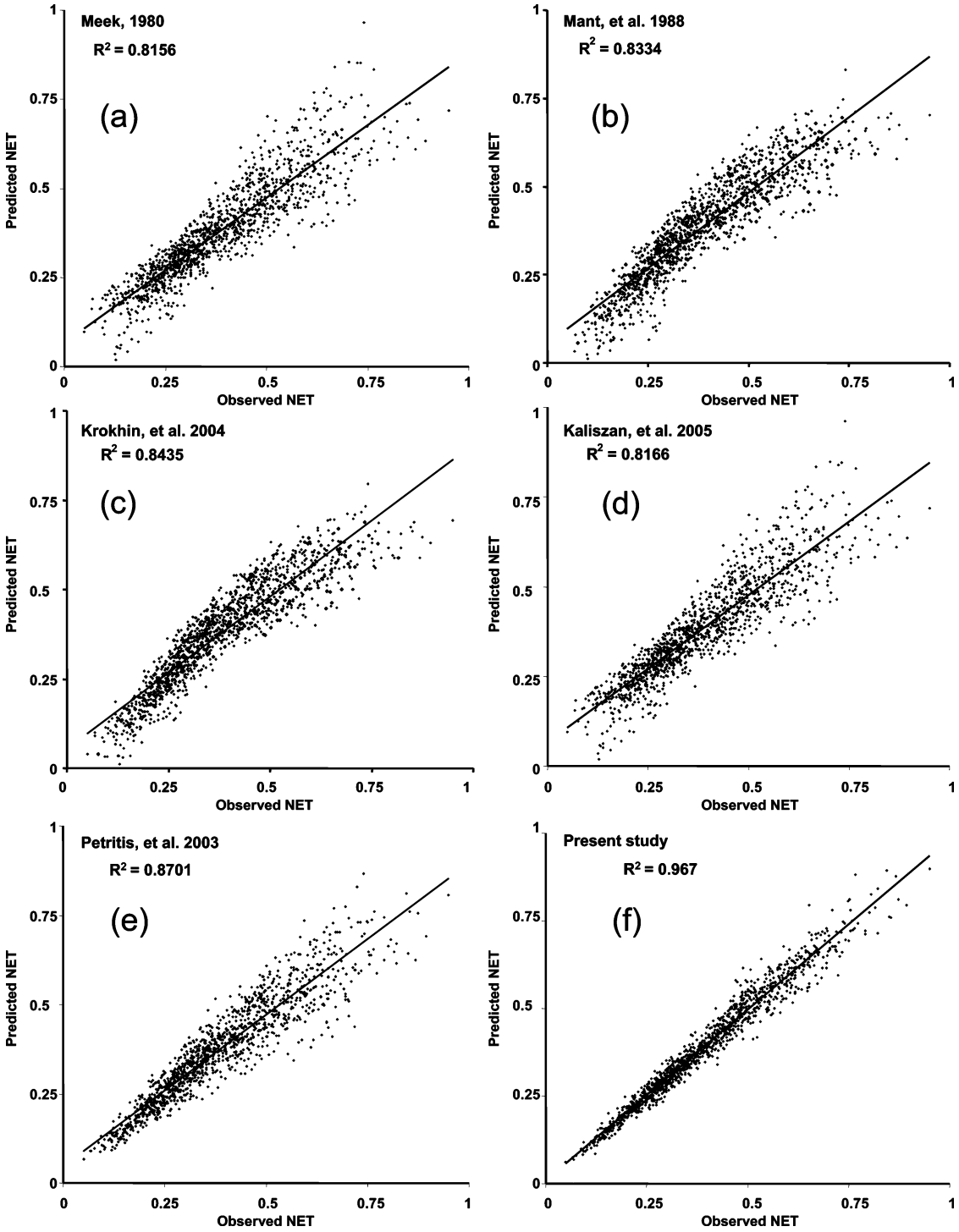
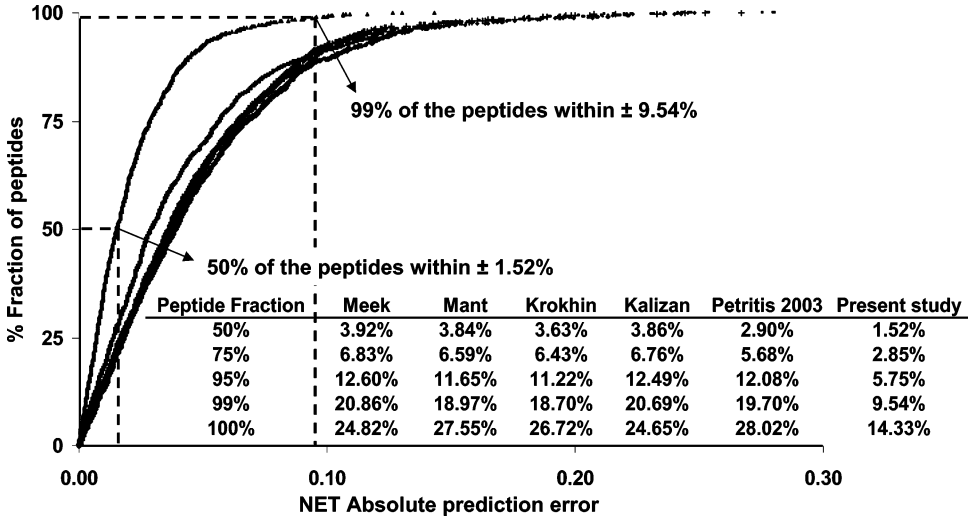


Figure 4. Comparison of peptide retention time prediction for 6 different models. The diagrams show the predicted vs observed normalized elution time correlations of each method for the 1303 confident peptide identifications of the testing set.

predictors, we were unable to reproduce the study by Liu et al.,[39] as it used software unavailable to us for calculating peptide constitutional and topological descriptors. In addition, the study by Kaliszan et al.[40] used HyperChem software to calculate the peptide hydrophobicity *c* log *P* and molecular volume VDW_{Vol}. We generated these values by using alternative published algorithms,[91,92] and our calculated values were similar to the Kaliszan values. The optimization for each model was accomplished by using genetic algorithms, which continued to optimize the variables in the algorithms to our data until the errors converged. A comprehensive table of the observed and predicted elution times obtained from the different predictors for the 1303 peptides of the testing test are provided as Supporting Information.

Figure 4 shows the correlation coefficients between observed and predicted NETs from five previously reported models and the 1052-24-1 ANN model developed in this study. All previously

(91) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441−445.
(92) Tao, P.; Wang, R. X.; Lai, L. H. *J. Mol. Model.* **1999**, *5*, 189−195.



Peptide Fraction	Meek	Mant	Krokhin	Kalizan	Petritis 2003	Present study
50%	3.92%	3.84%	3.63%	3.86%	2.90%	1.52%
75%	6.83%	6.59%	6.43%	6.76%	5.68%	2.85%
95%	12.60%	11.65%	11.22%	12.49%	12.08%	5.75%
99%	20.86%	18.97%	18.70%	20.69%	19.70%	9.54%
100%	24.82%	27.55%	26.72%	24.65%	28.02%	14.33%

Figure 5. Comparison of peptide retention time prediction for 6 different models. The diagram shows the prediction error distributions for 1303 confidently identified peptides. The method described in this study provides approximately 2-fold better predictions than previously described methods. Key: (▲) Present study, (−) Petritis et al. 2003,[43] (+) Krokhin et al. 2004,[38] (■) Mant et al. 1988,[13] (●) Kaliszan et al. 2005,[40] and (◆) Meek 1980.[10]

described peptide elution time predictors yielded inferior correlations compared with the ANN-based elution time predictors. The best model developed by other workers[38,90] provides a correlation coefficient of 0.8435. In general, all of these models performed similarly, with correlation coefficients varying from 0.8156 to 0.8435. This similarity in performance can be expected as all the models were based on the retention coefficient approach.[10] The Kaliszan et al.[40] model provided only small improvements over the Meek model;[10] as in our case, the genetic algorithm underweighted the *c* log *P* and VDW_{Vol} parameters. As expected, all models gave better correlations than the original Meek model.[10]

Figure 5 shows the prediction error distribution of all the prediction models. Note that the current model has 50% of the peptides within ±1.52% error and performs ~2-fold better than both our previously described model and any other described model. For all other models, 99% of the peptides were within ~±20% of their predicted values, while for our current model, 99% of the peptides were under ±10% of their predicted values, confirming the present model demonstrates a 2-fold improvement over all previously published models. Figures 4 and 5 indicate that this is the first major improvement in making more accurate peptide elution time predictions since the original work by Meek.

It must be noted here that the disadvantage of the present model over previously developed predictors is the large number of peptide identifications needed to train it. The generation of such a large training set is time-consuming and could limit its use from other groups that would like to reproduce the present predictor for different chromatographic conditions. However, to some extent it is possible to adapt the present model to separations with modest changes (e.g., gradient shape). We note that the development of improved alignment algorithms for LC−MS data provides a likely basis for effective alignment of data sets from different chromatographic systems (e.g., using different gradient shapes). Preliminary results[93] show that, by using a transformation function, it should also be possible to transform peptide elution time predictions from our current chromatographic system to other chromatographic systems without losing much predictive capability. This development, if validated, will facilitate broader application of the present model without the necessity of adopting all of the present chromatographic conditions or acquiring the large data sets of peptide identifications needed to develop the present predictive capability.

CONCLUSIONS

In this study, an improved ANN-based peptide retention time predictor was developed that provides an average error of 1.5%. Most of the improvement arises from incorporation of peptide sequence information into the model as opposed to simply amino acid composition. Moreover, the peptide length and hydrophobic moment provided additional small improvements in the model's prediction capability. Predictor encoding was limited to 50 amino acid residues since most present MS/MS data are limited to this regime. In addition to the 20 proteinogenic amino acids, the present model was trained to predict the retention time of peptides that contain alkylated cysteines. Unlike any of the previously developed predictors, this model is now able to accurately predict the retention times of both isobar and isomer peptides. Such capability allows more confident identification of isomeric/isobaric peptides otherwise indistinguishable by accurate mass measurements.

The development of the present predictive capability was enabled by the availability of large quantities of data accumulated over the years, using identical chromatographic conditions, and providing an extremely large set of confident peptide identifications. Approximately 346 000 peptides were used to train the ANN predictor. In addition, the development of a new generation of ion trap instruments from several manufacturers that offer faster cycle times and better sensitivities helped in accumulating sufficient training data. To the best of our knowledge, this is the first time that such massive quantities of proteomic data have been used for the development of a peptide retention time predictor. Due to the large amount of data that needed to be normalized, we revised the normalization procedure to an independent linear

(93) Jaitly, N.; Monroe, M. E.; Petyuk, V.; Clauss, T. R. W.; Adkins, J. N.; Smith, R. D. *Anal. Chem.*, in press.

regression for each analysis. In a comparison to previously reported models, our model provided ~2-fold improvement.

Finally, preliminary results indicate that it should also be possible to transform peptide elution time predictions from the current chromatographic system to other similar chromatographic conditions without losing much predictive capability. This development will facilitate broader application of the present model without the necessity of adopting all of the present chromatographic conditions or acquiring the large data sets of peptide identifications needed to develop the present predictive capability. We also plan to explore the use of ANNs for predicting the elution times of peptides with posttranslational modifications. This capability would allow the implementation of targeted experiments; that is, the expected mass of the modified peptide (if detected in the predicted elution window) would be added to the inclusion list of the masses to be selected for fragmentation. We further plan to apply the ANN approach to predict the elution time of peptides separated by ion-exchange chromatography for further quality assurance. This will add another dimension of confidence and will be especially useful for research groups that use on-line (e.g., MudPIT) or off-line strong cation-exchange columns for peptide separation/fractionation.

ACKNOWLEDGMENT

This work was supported by the NIH National Center for Research Resources (RR18522). Peptide identifications originated from projects funded by the National Institute of Allergy and Infectious Diseases (NIH/DHHS through interagency agreement Y1-AI-4894-01), the Genomics:GtL Program and the Chemical and Biological National Security Program (Office of Biological and Environmental Research, U.S. Department of Energy), and the Biological Countermeasures Program (Department of Homeland Security). The work was performed in the Environmental Molecular Sciences Laboratory, a U.S. Department of Energy (DOE) national scientific user facility located at the Pacific Northwest National Laboratory (PNNL) in Richland, WA. PNNL is a multiprogram national laboratory operated by Battelle for the DOE under Contract DE-AC05-76RL01830. Work by B.Y. and Y.X. was supported in part by National Science Foundation (NSF/DBI-0354771, NSF/ITR-IIS-0407204) and by the DOE Genomes:GtL Program (Carbon Sequestration in Synechococcus sp: From Molecular Machines to Hierarchical Modeling project).

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review January 19, 2006. Accepted April 17, 2006.

AC060143P