research articles Journal of proteome research

Prediction of Liquid Chromatographic Retention Times of Peptides Generated by Protease Digestion of the *Escherichia coli* Proteome Using Artificial Neural Networks

Kosaku Shinoda,[†,‡] Masahiro Sugimoto,[‡,§] Nozomu Yachie,[‡] Naoyuki Sugiyama,[†] Takeshi Masuda,[‡] Martin Robert,[‡] Tomoyoshi Soga,[*,†,‡] and Masaru Tomita[†,‡]

Human Metabolome Technologies, Inc., Tsuruoka, Yamagata 997-0052, Japan, Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017 Japan, and Bioinformatics Department, Mitsubishi Space Software Company Ltd., Amagasaki, Hyogo, 661-0001, Japan

Received May 1, 2006

We developed a computational method to predict the retention times of peptides in HPLC using artificial neural networks (ANN). We performed stepwise multiple linear regressions and selected for ANN input amino acids that significantly affected the LC retention time. Unlike conventional linear models, the trained ANN accurately predicted the retention time of peptides containing up to 50 amino acid residues. In 834 peptides, there was a strong correlation ($R^2 = 0.928$) between measured and predicted retention times. We demonstrated the utility of our method by the prediction of the retention time of 121 273 peptides resulting from LysC-digestion of the *Escherichia coli* proteome. Our approach is useful for the proteome-wide characterization of peptides and the identification of unknown peptide peaks obtained in proteome analysis.

Keywords: liquid chromatography • retention time prediction • stepwise multiple linear regression • artificial neural networks • peptide identification

Introduction

Liquid chromatography mass spectrometry (LC−MS) is a powerful tool for the separation and identification of peptides in proteomics studies. However, when chromatographic peaks are very small due to low peptide abundance or poor ionization in MS, it is difficult to obtain a clear MS/MS spectrum. As a consequence, most peptides cannot be identified, and proteome coverage remains limited. As the chromatographic retention times of peptides depend on their chemical composition, their retention time complements the information provided by MS and enhances their identification. The prediction of the chromatographic behavior of peptides in reversed-[1−7] and normal-phase chromatography[8,9] has been described. One approach uses a linear regression model[1−5,8,9] to estimate the retention time of peptides from their amino acid sequence. However, this method is limited because it cannot be applied to peptides longer than 15−20 amino acids.[7] Another prediction method is based on artificial neural networks (ANN). It applies a machine-learning algorithm for solving nonlinear problems[7,10−20] and has been used to model the quantitative structure−retention relationships (QSRR) of various analytes in liquid chromatography[10−12] and for predicting migration times in capillary electrophoresis.[13−16] A previously reported ANN method for the prediction of the liquid chromatographic retention times of peptides utilized a multiple-layer architecture consisting of 20 input nodes that corresponded with the 20 different amino acids.[7]

There are several studies on the prediction of LC retention time of tryptic peptides for protein identification.[2,3,21−23] However, the large-scale prediction of LC retention time of LysC-digested peptides has not been reported. LysC endopeptidase is a highly efficient and specific protease that is widely used in quantitative proteomics because it quantitatively digests proteins at the C-terminal side of all lysines. The use of LysC produces longer peptides that need a better prediction model than existing ones. We trained an ANN using experimentally derived LC retention times of peptides obtained from a set of recombinant proteins and developed a solid retention time prediction tool that uses only amino acid sequence information. We then used this ANN to predict the retention time of all peptides generated by LysC digestion of the *Escherichia coli* (*E. coli*) K12 proteome.

Experimental Section

Preparation of LysC-Digested Peptides. The peptides used to train the ANN were produced by LysC digestion of recombinant proteins. Recombinant histidine-tagged proteins were produced in *E. coli* expressing selected single open reading frames (ORFs) obtained from the ASKA library.[24] Cultures were grown in LB medium for 3 h at 37 °C, and then 1 mM IPTG was added to induce the expression of recombinant proteins.

[*] To whom correspondence should be addressed. Professor Tomoyoshi Soga, Ph.D., Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017, Japan. E-mail, soga@sfc.keio.ac.jp; tel, +81-235-29-0528; fax, +81-235-29-0530.

[†] Human Metabolome Technologies, Inc.

[‡] Keio University.

[§] Mitsubishi Space Software Co. Ltd.

10.1021/pr0602038 CCC: $33.50

Published on Web 11/02/2006

Cells, harvested by 10-min centrifugation at 5000 rpm, were suspended in sodium phosphate buffer (50 mM sodium phosphate and 300 mM NaCl, pH 7.0), disrupted by sonication on ice, and freeze-thawed. Protein extracts were collected by 20-min centrifugation at 8000 rpm at 4 °C, and the supernatant was applied to a column containing TALON Metal Affinity resin (Clontech, CA). After column washing, His-tagged proteins were eluted in breakage buffer containing 200 mM imidazole. Respective proteins (total amount 12.5 μg), dissolved in 20 μL of 0.1 M Tris-HCl (pH 9.0) containing 6 M guanidine-HCl, were reduced with 1 μL of reduction buffer (1 μg/μL dithiothreitol in water), alkylated with 1 μL of alkylation buffer (5 μg/μL iodoacetic acid in water), diluted with Milli-Q water, and enzymatically digested overnight at room temperature by adding 10 AU/mL LysC endopeptidase. The reaction was terminated by adding 5 μL of 10% trifluoroacetic acid, and then 1 μL of 125 μM testosterone was added to the reaction mixture, as an internal standard.

LC−MS/MS Experiments. The peptide mixture was injected into ZORBAX SB-C18 reversed-phase columns (2.1 mm × 50 mm, 3.5 μm, Agilent Technologies, Palo Alto, CA) packed with 1.8-μm particles equilibrated with 0.2% formic acid. After 5-min of column conditioning with the same solvent, peptides were eluted at a flow rate of 0.2 mL/min using a two-step gradient elution; the acetonitrile concentration was first linearly increased from 1 to 40% over 50 min and then raised to 50% for 5 min. The eluent was transferred to a hybrid quadrupole time-of-flight mass spectrometer (QSTAR; Applied Biosystems, CA). The ion spray voltage was set at 5.5 kV in the positive mode. The scanning range of Q1 and Q2 was m/z 200−1000 for 0.5 s/scan and m/z 100−1000 for 0.5 s/scan, respectively. The LC−MS/MS chromatograms were manually interpreted and deconvoluted, and peptide peak retention times were determined. The retention time of all peptides was normalized to the retention time of testosterone (internal standard). All LC−MS steps were controlled by the Applied Biosystems Analyst software.

Table 1. The Results of Stepwise Multiple Linear Regressions (SMLR)[a]

amino acid	RC	*F*-value	*P* (prob > *F*)	hydrophobicity
Leucine	0.46	471.45	0.0	−1.82
Isoleucine	0.36	263.94	0.0	−1.82
Phenylalanine	0.38	256.98	0.0	−2.27
Arginine	−0.34	128.30	0.0	3.95
Valine	0.25	126.59	0.0	−1.3
Tryptophan	0.27	102.77	0.0	−2.13
Methionine	0.28	56.77	0.0	−0.96
Histidine	−0.32	56.54	0.0	0.64
Tyrosine	0.25	51.08	0.0	−1.47
Alanine	0.14	15.84	0.0	−0.39
Glycine	−0.08	8.06	0.005	0
Asparagine	−0.06	6.95	0.009	1.91
Proline	0.05	6.60	0.010	−0.99
Lysine	−0.06	6.48	0.011	2.77
Glutamic acid	0.06	5.17	0.023	2.91
Cysteine	0.08	4.57	0.033	−0.25
Serine	−	3.71	0.055	1.24
Threonine	−	1.44	0.231	1
Asparatic acid	−	1.01	0.316	3.81
Glutamine	−	0.18	0.672	1.3

[a] Statistical and hydrophobicity parameters for the 20 amino acids are listed. RC indicates the regression coefficient obtained from SMLR. Relative hydrophobicities were from Creighton[28] and were calculated from the hydrophobicities of the individual groups that make up each side chain.[29] The larger the negative value, the stronger the hydrophobicity.

Artificial Neural Networks. The development of the ANN model was based on the assumption that the LC retention time of peptides depends on their amino acid composition. Therefore, we used the number of residues of each amino acid in the peptides as descriptors and employed them as inputs for the ANN. Neither a positive nor a negative correlation was observed ($P < 0.01$) between pairs of the amino acid residues. We applied stepwise multiple linear regression (SMLR) to select significant descriptors among the 20 amino acids. Forward and backward selection procedures were combined using the number of each amino acid contained in the peptides as the independent variable and the LC retention time as the dependent variable. Stepwise selection ($P < 0.05$) eliminated 4 of the 20 original descriptors. The ANN was trained with the 16 remaining descriptors as inputs and normalized retention time (NRT) as outputs. The ANN used in this study was a three-layer architecture (input, output, and hidden layer) model with back-propagation learning. A sigmoid function was applied for each node in the ANN. We chose a three-layer architecture because it could approximate any functions.[25] The values of ANN inputs and outputs were normalized to a numerical value between 0.1 and 0.9. We applied a two-deep cross-validation method[26] to assess the prediction and generalization ability of the ANN. A total of 834 peptides was randomly divided into the cross-validation set (including 80% in the training and 10% in the test set) and the final test set (the remaining 10%). The final test set was not used under cross-validation loops. The two-deep cross-validation was repeated 10 times until all peptides had been selected as a member of the final test set. For each validation, training was performed 20 times with different initial weights that corresponded to the numerical values between the ANN nodes in order to avoid local minima. Therefore, for each combination in the cross-validation, ANNs were trained for a total of 200 times with different initial weight values. To reduce the unnecessarily large parameters (weights) among nodes, the optimization function was formed by the sum of squared residuals plus the over-fitting penalty times the sum of squares of the parameter estimates (pruning method). Each training was continued until epoch (iteration) reached 100 or improvement of the optimization function fell below a learning convergence criterion (see ref 27 for detail). After iterative cross-validation, the ANN with the highest correlation coefficient was selected for subsequent proteome-wide peptide LC retention time prediction. This computational portion of our work was performed using JMP software version 5.1.2 (SAS Institute, Cary, NC).

Results and Discussion

ANN Construction for LC Retention Time Prediction. The first step in building the ANN for retention time prediction was to select descriptors of chromatographic behavior. Table 1 shows the results of SMLR to select such descriptors. Regression coefficients, *F*-values, *P*-values derived from the *F*-values, and the relative hydrophobicity of each amino acid are shown. When the *P*-value, suggestive of the reliability of the *F*-value, was less than 0.05, that amino acid was added to the SMLR model. With this procedure, 16 of the original 20 amino acids were selected. The excluded amino acids were serine, threonine, asparatic acid, and glutamine. Threonine and serine are characterized by neutral properties with near-zero relative hydrophobicity values of 1 and 1.24, respectively. Glutamine has a low-level hydrophilic property (1.3). On the basis of these results, we postulated that amino acids with neutral properties have negligible effects on the retention time, and they were eliminated from the descriptors for ANN through stepwise variable selection. Among the excluded amino acids, asparatic acid generally has strong hydrophilic property; however, in the mobile phase we used (0.2% formic acid, pH = 2.4), the side chain carboxylic group ($\mathrm{p}K_a = 4.5$)[30] is almost (>99%, estimated using the equation: $\mathrm{pH} = \mathrm{p}K_a + \log[\mathrm{A}^-]/[\mathrm{AH}]$) undissociated; thus, it can be treated as neutral and has less effect on LC retention time in spite of its high hydrophilicity. The same is true of glutamic acid whose side chain $\mathrm{p}K_a$ is 4.6.[30] Glutamic acid was selected for the descriptors; however, it showed the second-lowest *F*-value (5.17) among the 16 selected descriptors and a *P*-value of 0.023, indicating relatively limited effects on LC retention time. Overall, in spite of their high hydrophilicity, glutamic acid and asparatic acid had low effects on LC retention time because of the acidic buffer condition. The difference in the effects of asparatic and glutamic acid on LC retention time may be explicable by differences in the length of their carbon chains and their contribution to hydrophobicity.

Petritis et al.[7] previously evaluated the degree of connection between ANN inputs and outputs based on trained ANN weights. We also analyzed the relationship between inputs and outputs using both the regression coefficients and the *F*-values of the SMLR model. Leucine, isoleucine, and phenylalanine had the highest regression coefficient (mean 0.4); they are strongly hydrophobic. In contrast, arginine, which manifested the highest hydrophilicity index among the 20 amino acids, had the lowest (−0.34) regression coefficient. These results suggest a positive correlation between the number of hydrophobic amino acids and the LC retention time in reversed-phase columns. As a whole, among selected descriptors, highly hydrophobic or hydrophilic amino acids tended to have high *F*-values, indicating a significant effect on the LC retention time.

A hidden layer with too few nodes may not model the data sufficiently, while a hidden layer with too many nodes may over-fit the data and result in the loss of generalization ability. We tried the hidden nodes ranging from 2 to 10, and the retention time response curves of each input variable, constituting an approximate function from sampled values, were used as the criteria for determining the number of hidden layers; that is, we added hidden nodes until the response curves were not too flexible or nonlinear. Consequently, we adopted 4 nodes in the hidden layer. As shown in Figure 1, our ANN had a 16-4-1 architecture. Other parameters for ANN training (training ratio, momentum, and random numbers for initial ANN weights) were determined empirically.

Assessment of Prediction Ability of the ANN. To test the validity of our ANN for retention time prediction, we compared computationally and experimentally obtained results. Figure 2 is a global comparison between predicted and measured NRT for 834 peptides through two-deep cross-validations (see Experimental Section for details). Overall, our results were very good; the correlation coefficient was 0.928. Experimentally measured and predicted NRTs for a randomly chosen subset of the data are listed in Table 2. The complete list is available in Supporting Information Table 1. Of the 834 NRT values, 546 (65.5%) were within a range of 0.4−0.7, forming a normal distribution (mean 0.55; variance 0.03). The mean prediction error was 0.0322 ± 3.9% (relative standard deviation, RSD). The 705 peptides (84.5%) that eluted between 0.4 and 0.9 NRT had significantly ($P < 0.01$, *t*-test) smaller prediction errors (mean 0.025) than the 129 peptides (15.5%) that eluted between 0.1 and 0.4 NRT (mean 0.033), the latter representing the majority of outliers. We postulate that this retention-time-dependent difference in prediction accuracy may be due to higher amounts of salt eluting during the earlier part of LC separation. Norbeck et al. showed that adopting LC-retention-time-based peptide screening with a constraint of ±0.05 NRT (normalized between 0 and 1) in proteome-wide peptide determination significantly increased the uniqueness of peptides.[23] The calculated mean prediction error of our ANN by two-deep cross-validations was 0.0322 NRT (normalized between 0.1 and 0.9) as mentioned above. Therefore, we suggest that our ANN facilitates the identification of peptides.

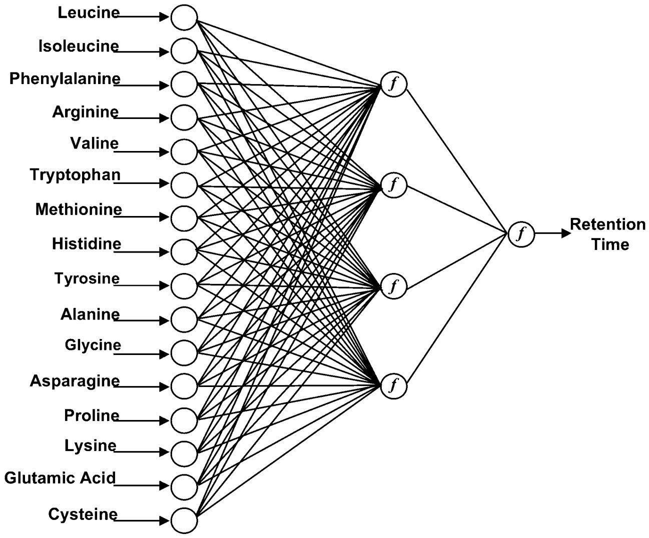

Figure 1. Architecture (16-4-1) of the ANN used in this study. The name of the amino acid reflects the normalized number of each amino acid contained in the peptide.

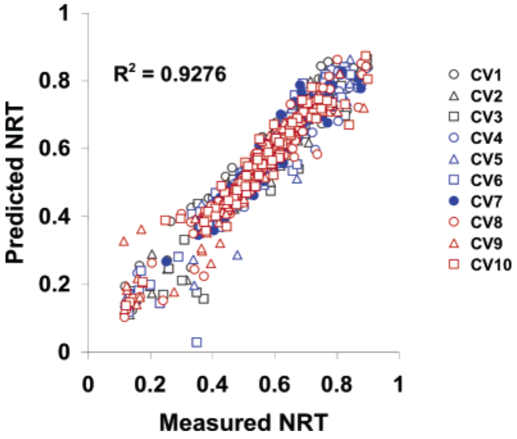

Figure 2. Scattergram of the correlation between experimentally measured and predicted normalized LC retention times for all 834 peptides through 10-fold two-deep cross-validations (CV1−10).

To assess confidence intervals of predicted LC retention times, we prepared LC retention time data doped with artificial normally distributed noises. Prediction with the noisy data was conducted 10 times using different noise levels. Figure 3 summarizes the relationship between the noise level and prediction accuracy. The coefficient of determination (R^2) and the sum of square errors (SSE) of LC retention time prediction are shown. For comparison, we analyzed day-to-day (total of 6 days) LC retention time fluctuation (i.e., RSD) of 55 different peptides under the same LC conditions. Among these 55 peptides, the RSDs of LC retention time varied from 0.57 to

Table 2. Experimentally Measured and Predicted NRTs of the Peptide Validation Data Set[a]

	NRT			NRT	
peptide seq.	meas.	pred.	peptide seq.	meas.	pred.
ASAQLETIK	0.431	0.437	FGEIEEVELGRIQK	0.566	0.601
YFPDATILALTTNEK	0.671	0.666	GDPVYLK	0.448	0.454
MVEVNACLK	0.504	0.496	IVVFNNSVLGFVAMEMK	0.774	0.774
QHEFSHATGELTALLSAIK	0.674	0.623	NGVEERK	0.127	0.090
LDEAGVRMIGPNCPGVITPGECK	0.586	0.611	TCPK	0.121	0.133
VNEDLGLLSEEK	0.547	0.561	FFPAEANGGVK	0.497	0.509
NHLNMHFVSNVDGTHIAEVL K	0.587	0.621	GFSGEDATPALEGADVVLISAGVARK	0.679	0.683
GFTSEITVTSNGK	0.500	0.480	ERVTEAFK	0.410	0.391
DAASFAPLHNPAHLIGIEEALK	0.651	0.661	IIQIDINPASIGAHSK	0.596	0.584
LGELLEALK	0.608	0.597	NYDPRATVMRETCHEVLK	0.527	0.534
RAMDVYCHRLAK	0.440	0.443	IYAYLSRGLCGR	0.545	0.541
ITDAYAENPQIANLLLAPYF K	0.745	0.754	LLPWIDGLLDAGEK	0.753	0.737
QEAAPAAAPAPAAGVK	0.438	0.460	LLAWLETLK	0.679	0.682
QVIDASHAEGK	0.371	0.372	IIASVAEK	0.420	0.455
PLSPETWQHLK	0.529	0.530	HLINK	0.171	0.194
SIREAGVQEADFLANVDK	0.581	0.596	VLGVK	0.386	0.400
VGIVSRSGTLTYEAVK	0.526	0.562	DSVSYGVVK	0.452	0.436
IAAGDTSNLGDTSTLADPGVVEK	0.550	0.590	NGLACITPISALNQPGK	0.622	0.600
AMQEVLQQFAHVK	0.599	0.575	ELGTK	0.141	0.140
LASAK	0.123	0.172	ARHLVDLYQQQGVEK	0.477	0.503
GVNLPGVSIALPALAEK	0.698	0.656	FLCIAK	0.510	0.493
VTGQALTVNEK	0.431	0.449	AGQTFTFTTDK	0.501	0.508
VIVVTDGERILGLGDQGIGGMGIPIGK	0.700	0.682	LSPTLAMYRAK	0.505	0.523
RALIVTDRFLFNNGYADQITSVLK	0.687	0.755	LPYITFPEGSEEHTYLHAQRQK	0.572	0.581
APVIVQFSNGGASFIAGK	0.635	0.648	LIQLMNETVDGDYQTFK	0.630	0.667
TDITELEAFRK	0.538	0.525	NINPTQTAAWQALQK	0.575	0.561
TRNAIIFSPHPRAK	0.436	0.465	VMWGSRWDELLRK	0.608	0.626
DWGYQLAREEFGGELIDGGPWLK	0.729	0.742	AVLNRGVSVVVLPGDVALK	0.629	0.672
LMPEFWQFPTVSMGLGPIGAIYQAK	0.812	0.827	EYLPASYHEGSK	0.443	0.454
ESAPAAAAPAAQPALAARSEK	0.468	0.468	GVHEGHVAAEVIAGK	0.444	0.427
ELLEDPTRLLLDVGLCGR	0.733	0.731	RGLCGR	0.242	0.209
SVLCIGGSWLVPADALEAGDYDRITK	0.723	0.736	MRTLGTAACPPYHIAFVIGG TSAETNLK	0.642	0.687
LVSWYDNETGYSNK	0.540	0.556	MVGGVTPGK	0.410	0.426
VHPNDDVNK	0.264	0.294	NLDDK	0.136	0.135
RPFK	0.144	0.158	EYAPAEDPGVVSVSEIYQYYK	0.664	0.638
MVPCDFIAPAITHNPLSDHH QK	0.629	0.622	HQFAQSLNYEIAK	0.518	0.511
APNSPVAGRATVFIFPDLNTGNTTYK	0.660	0.653	HVTWRGEAIPGSLFDFALYFFHNYQALLAK	0.880	0.815
RMQAAGAQAYLVNTGWNGTGK	0.555	0.564	EDGIYVTMEGK	0.513	0.509
EGTRPAVVIPTNEELVIAQDASRLTAGLCGR	0.720	0.700	IINELEGIFEGAGWNVIK	0.756	0.782
IAFGCDHVGFILK	0.632	0.651	LRIMFPMIISVEEVRALRK	0.698	0.728
ESGLLGLTEVTSDCRYVEDNYATK	0.648	0.623	ANSGHPGAPMGMADIAEVLWNDFLK	0.826	0.740
QLIPQLK	0.539	0.502	TAGVIGTGK	0.374	0.387
GVVPQLVK	0.524	0.475	YIPEELRK	0.440	0.458
HFSTTPAEK	0.355	0.357	LMAELEGLCGR	0.583	0.609
LNERHYGALQGLNK	0.454	0.453	ALLSMAIRAAK	0.549	0.503
IAEAAVVGIPHNIK	0.537	0.532	FCQEEDK	0.344	0.358
YYDELPTEGNEHGQAFRDVE LEK	0.552	0.572	IAELAGFSVPENTK	0.581	0.583
TTLSTDPK	0.371	0.361	LELPSLQDFGALLEEQSK	0.762	0.729
MLNK	0.163	0.204	ERAADVRDIGK	0.382	0.352
AIEGTDRSSLRILGSVGEPINPEAWEWYWK	0.727	0.758	DWRGGRGASQNIIPSSTGAAK	0.490	0.468
QTYCGPIGAEYMHITSTEEK	0.581	0.558	GFDSPREFYVGRLTEGIATL GAAFYPK	0.747	0.708
PARVVMEK	0.369	0.412	AVAAVNGPIAQALIGK	0.622	0.593
DSILEAIDAGIK	0.662	0.588	LAVNFGAEILK	0.634	0.606
FSATFDDQMLVDYSK	0.625	0.651	LNAK	0.124	0.149
VIPSIAYTEPEVAWVGLTEK	0.708	0.723	AEAPAAAPAAK	0.365	0.367
TRTQQIEELQK	0.422	0.456	FNLMLETK	0.577	0.574
IFMGLATIFLERDLALIEINPLVITK	0.897	0.868	EAAPAAAPAAAAAK	0.404	0.432
VYYDLLEQRRK	0.497	0.494	DFEDAVEK	0.451	0.443
QRSLYIPYAGPVLLEFPLLNK	0.767	0.742	AGVYDK	0.199	0.203
HFDEMK	0.330	0.319	DVTIADLFAK	0.658	0.590
AGRGLCGR	0.310	0.320	HYFDPK	0.387	0.355
YVALQFLRNSDIAAK	0.606	0.619	PGVITGDDVQK	0.438	0.474
YIALRCAGFNNVDLDAAK	0.595	0.628	FPLPVEVIPMARSAVARQLVK	0.703	0.693
RYYLGNADEIAAK	0.498	0.492	LAPSLTLGCGSWGGNSISENVGPK	0.631	0.643
NRITEETLAK	0.418	0.410	LQTLGLTQGTVVTISAEGED EQK	0.607	0.647
LLAYNCSPSFNWQK	0.609	0.617	IYQLK	0.426	0.436
HFAALSTNAK	0.415	0.429	VLPAVAMLEERAK	0.617	0.573
QMNDEIHQNLVGVSNHRTLEFAK	0.545	0.581	GDVLNYDEVMERMDHFMDWLAK	0.783	0.734
LDNLVFVINCNLQRLDGPVTGNGK	0.730	0.706	EEAHGAPLGEEEVALARQK	0.497	0.498
EGVFHTEWLDGLCGR	0.647	0.653	GIANSILIK	0.552	0.515
TLGEFIVEK	0.556	0.565	IGVAMLRILK	0.610	0.610

Table 2. (Continued)

peptide seq.	NRT meas.	NRT pred.	peptide seq.	NRT meas.	NRT pred.
LSEDAFDDQCTGANPRYPLISELK	0.629	0.651	LRGSVNPECTLAQLGAAK	0.583	0.569
RISTVPEAVEMQSRVAK	0.500	0.512	DITLAMDCAASEFYK	0.661	0.657
NDSFRLMGFGHRVYK	0.532	0.542	AAYSSGK	0.133	0.149
RLLTTCNIPVPSDVRVATEFSETAPATLK	0.656	0.678	GMNTAVGDEGGYAPNLGSNAEALAVIAEAVK	0.727	0.719

[a] For the complete list, see Supporting Information Table 1.

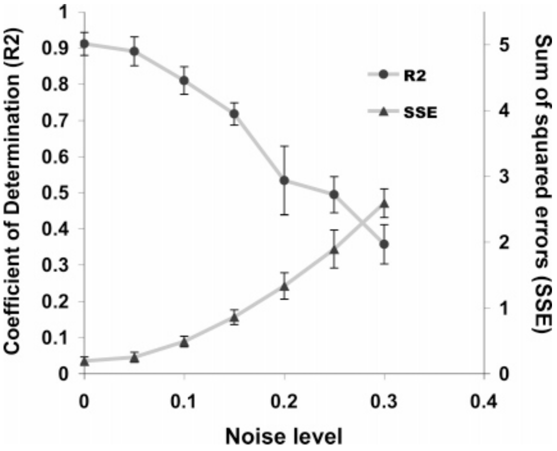

Figure 3. Relationship between prediction accuracy and artificial noise level of LC retention time ($n = 10$). The coefficient of determination and the sum of squared errors are shown. The *X*-axis indicates the noise level (with average of 0), which corresponds to the standard deviation of the normal distribution function used for noise generation.

1.21% (mean: 0.79). This result agrees with other published data on LC retention time reproducibility.[31] With this level of error (noise level < 0.05), our ANN achieved R^2 of >0.89 and SSE of <0.25, and thus, we can conclude that prediction accuracy did not drastically deteriorate as a result of the errors. These results demonstrate the utility of our ANN in experimental practice.

Retention Time Prediction for *E. coli* Proteome-Derived Peptides. We next used the trained ANN to predict the retention time of all peptides generated by digestion of the *E. coli* K12 proteome. We chose *E. coli* as an appropriate model because its proteome manifests relatively few post-translational protein modifications and because the protein-coding sequences are well-characterized. To construct the complete peptide data set, we used annotated GenBank data (National Center for Biotechnology Information, Jan 12, 2004). All *E. coli* protein sequences in the GenBank database were computationally digested based on the rule of LysC endopeptidase cleavage. The number of missed cleavages was set at ≤1. Eliminating duplicated peptide sequences, the resulting dataset contained 121 273 potential peptides. LC retention times were predicted for the whole data set using the trained ANN. The complete data sets, including the amino acid sequences and the predicted retention time of all peptides, are listed in Supporting Information Tables 2–4. Many of the retention time values (28 677 of 121 273, 23.6%) were predicted to fall in the relatively late, 25–30-min range. Peptides predicted to elute within this range were longer; their mean molecular weight (MW) was 5357 compared to earlier-eluting peptides whose mean MW was 2540. This link between peptide length and elution time is well-known, and our predictions agree with this. Although others have reported methods for predicting the LC retention time of tryptic/synthetic peptides and its application to protein identification,[2,3,21–23] to our knowledge, there are no published proteome-wide retention time predictions for the larger LysC-digested peptides.

Our method can be used for predicting the elution order within a given peptide data set. Prediction of the elution order and comparison with experimentally measured results make it possible to rank peptide candidates based on their LC retention times and to increase the probability of correct identification. Even if the ranking is not completely matched, this method can still determine the priority for peptide candidates. From this perspective, we estimated the error rate in the elution order prediction using the following procedure. Elution order prediction can be generalized down to a prediction of the anteroposterior relations of each peptide pair. On the basis of this generalization, we compared predicted and experimental 834 × 833 anteroposterior relations of peptide retention times within the validation data sets. We found that the mean prediction error was 5.8 ± 2.6 (SD)%. This result indicates that our ANN can predict the elution order of peptides with an error of <11%. Even in cases where peptides cannot be distinguished easily due to close *m*/*z* values, our method can facilitate peptide identification based on the predicted elution order. We suggest that the use of the elution order prediction described here, in addition to standard protein database search tools, can facilitate and improve peptide/protein identification.

Conclusions

We developed a new artificial neural network model for the prediction of peptide retention times in liquid chromatography. The NRTs of 834 peptides generated by LysC-digestion of recombinant proteins were experimentally measured by LC–MS/MS and employed for retention time prediction using only the amino acid composition. We demonstrated that, in most instances, our method facilitates the accurate prediction of the retention time of peptides. The LC retention times of longer peptides containing up to 50 amino acid residues were predicted with an error rate of <4%.

The utility of our method was demonstrated by predicting the retention time of 121 273 peptides generated from *in silico* digestion of the *E. coli* K12 proteome. The LC retention times prediction data will be useful to remove ambiguities during peptide identification. As our technique is directly applicable to any peptide data set, it is a powerful tool to determine the elution order of peptides generated by protease digestion of any proteome and will facilitate the identification of peptides in complex proteomes.

Acknowledgment. We thank Dr. Yasushi Ishihama, Dr. Natsumi Saito, and Yuji Kakazu for technical advice and support. This work was supported in part by a grant from the

Ministry of Education, Culture, Sports, Science and Technology (MEXT) (the 21st Century COE Program entitled "Understanding and Control of Life's Function via Systems Biology") and a grant from Grant-in-Aid of Creative Researches and Leading Project for Biosimulation from MEXT, as well as research funds from the Yamagata Prefectural Government and Tsuruoka City.

Supporting Information Available: Tables showing the experimentally measured and predicted NRTs of all 834 peptides (Supporting Information Table 1), and the complete data sets, including the amino acid sequences and the predicted retention time of all peptides (Supporting Information Tables 2−4). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Meek, J. L. *Proc. Natl. Acad. Sci. U.S.A.* **1980**, *77*, 1632−1636.

(2) Palmblad, M.; Ramström, M.; Markides, K. E.; Håkansson, P.; Bergquist, J. *Anal. Chem.* **2002**, *74*, 5826−5830.

(3) Palmblad, M.; Ramström, M.; Bailey, C. G.; McCutchen-Maloney, S. L.; Bergquist, J.; Zeller, L. C. *J. Chromatogr., B: Anal. Technol. Biomed. Life Sci.* **2004**, *803*, 131−135.

(4) Baczek, T.; Wiczling, P.; Marszałł, M.; Heyden, Y. V.; Kaliszan, R. *J. Proteome Res.* **2005**, *4*, 555−563.

(5) Kaliszan, R.; Baczek, T.; Cimochowska, A.; Juszczyk, P.; Wiśniewska, K.; Grzonka, Z. *Proteomics* **2005**, *5*, 409−415.

(6) Sanz-Nebot, V.; Toro, I.; Barbosa, J. *J. Chromatogr., A* **2001**, *933*, 45−56.

(7) Petritis, K.; Kangas, L. J.; Ferguson, P. L.; Anderson, G. A.; Pasa-Tolić, L.; Lipton, M. S.; Auberry, K. J.; Strittmatter, E. F.; Shen, Y.; Zhao, R.; Smith, R. D. *Anal. Chem.* **2003**, *75*, 1039−1048.

(8) Yoshida, T. *J. Chromatogr., A* **1998**, *808*, 105−112.

(9) Yoshida, T.; Okada, T. *J. Chromatogr., A* **1999**, *841*, 19−32.

(10) Agatonovic-Kustrin, S.; Zecevic, M.; Zivanovic, L. *J. Pharm. Biomed. Anal.* **1999**, *21*, 95−103.

(11) Tham, S. Y.; Agatonovic-Kustrin, S. *J. Pharm. Biomed. Anal.* **2002**, *28*, 581−590.

(12) Ruggieri, F.; D'Archivio, A. A.; Carlucci, G.; Mazzeo, P. *J. Chromatogr., A* **2005**, *1076*, 163−169.

(13) Jalali-Heravi, M.; Garkani-Nejad, Z. *J. Chromatogr., A* **2001**, *927*, 211−218.

(14) Jalali-Heravi, M.; Garkani-Nejad, Z. *J. Chromatogr., A* **2002**, *971*, 207−215.

(15) Malovaná, S.; Frías-García, S.; Havel, J. J. *Electrophoresis* **2002**, *23*, 1815−1821.

(16) Sugimoto, M.; Kikuchi, S.; Arita, M.; Soga, T.; Nishioka, T.; Tomita, M. *Anal. Chem.* **2005**, *77*, 78−84.

(17) Hammer, C. L.; Small, G. W.; Combs, R. J.; Knapp, R. B.; Kroutil, R. T. *Anal. Chem.* **2000**, *72*, 1680−1689.

(18) Jalali-Heravi, M.; Fatemi, M. H. *J. Chromatogr., A* **2000**, *897*, 227−235.

(19) Aires-de-Sousa, J.; Hemmer, M. C.; Gasteiger, J. *Anal. Chem.* **2002**, *74*, 80−90.

(20) Muzikár, M.; Havel, J.; Macka, M. *Electrophoresis* **2003**, *24*, 2252−2258.

(21) Strittmatter, E. F.; Kangas, L. J.; Petritis, K.; Mottaz, H. M.; Anderson, G. A.; Shen, Y.; Jacobs, J. M.; Camp, D. G., II; Smith, R. D. *J. Proteome Res.* **2004**, *3*, 760−769.

(22) Kawakami, T.; Tateishi, K.; Yamano, Y.; Ishikawa, T.; Kuroki, K.; Nishimura, T. *Proteomics* **2005**, *5*, 856−864.

(23) 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.

(24) Kitagawa, M.; Ara, T.; Arifuzzaman, M.; Ioka-Nakamichi, T.; Inamoto, E.; Toyonaga, H.; Mori, H. *DNA Res.* **2005**, *12*, 291−299.

(25) Funahashi, K. *Neural Networks* **1989**, *2*, 183−192.

(26) Jonathan, P.; Krzanowski, W. J.; McCarthy, W. V. *Stat. Computing* **2000**, *10*, 209−229.

(27) *JMP Statistics and Graphics Guide*; Release 6; SAS Institute Inc.: Cary, NC, 2005; pp 499−500.

(28) Creighton, T. E. *Proteins: Structures and Molecular Properties*, 2nd ed.; W. H. Freeman & Co., Publishers: New York, 1993; pp 153−155.

(29) Roseman, M. A. *J. Mol. Biol.* **1988**, *200*, 513−522.

(30) Tandford, C. *Adv. Protein Chem.* **1962**, *17*, 69−165.

(31) Murata, K.; Mano, N.; Asakawa, N.; Ishihama, Y. *J. Chromatogr., A* **2006**, *1123*, 47−52.

PR0602038