Large-Scale Prediction of Cationic Metabolite Identity and Migration Time in Capillary Electrophoresis Mass Spectrometry Using Artificial Neural Networks

ARTICLE *in* ANALYTICAL CHEMISTRY · FEBRUARY 2005

Impact Factor: 5.64 · DOI: 10.1021/ac048950g · Source: PubMed

Masahiro Sugimoto
Keio University
68 PUBLICATIONS **1,371** CITATIONS

Shinichi Kikuchi
Georgia Institute of Technology
20 PUBLICATIONS **476** CITATIONS

Takaaki Nishioka
Nara Institute of Science and Technology
173 PUBLICATIONS **7,646** CITATIONS

Masaru Tomita
Keio University
533 PUBLICATIONS **17,293** CITATIONS

Anal. Chem. **2005,** *77,* 78−84

Large-Scale Prediction of Cationic Metabolite Identity and Migration Time in Capillary Electrophoresis Mass Spectrometry Using Artificial Neural Networks

Masahiro Sugimoto,[†,‡] Shinichi Kikuchi,[†] Masanori Arita,[†,§,||] Tomoyoshi Soga,[*,†,⊥] Takaaki Nishioka,[†,#] and Masaru Tomita[†,⊥]

Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017, Japan, Mitsubishi Space Software Co. Ltd., Amagasaki, Hyogo 661-0001, Japan, Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa 277-8561, Japan, Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Aomi 2-43-17, Koto-ku, Tokyo 135-0064, Japan, Human Metabolome Technology Inc., Tsuruoka, Yamagata 997-0017, Japan, and Graduate School of Agricultural Sciences, Kyoto University, Kyoto 606-8502, Japan

We developed a computational technique to assist in the large-scale identification of charged metabolites. The electrophoretic mobility of metabolites in capillary electrophoresis−mass spectrometry (CE−MS) was predicted from their structure, using an ensemble of artificial neural networks (ANNs). Comparison between relative migration times of 241 various cations measured by CE−MS and predicted by a trained ANN ensemble produced a correlation coefficient of 0.931. When we used our technique to characterize all metabolites listed in the KEGG ligand database, the correct compounds among the top three candidates were predicted in 78.0% of cases. We suggest that this approach can be used for the prediction of the migration time of any cation and that it represents a powerful method for the identification of uncharacterized CE−MS peaks in metabolome analysis.

Metabolomics, the qualitative and quantitative analysis of all metabolites in a biological sample, has become a powerful new tool for gaining insights into cellular function. Since proteins and metabolites are the main effectors of phenotype, tracking metabolite changes under different conditions not only provides direct information on metabolic phenotypes but also complements gene expression and proteome analyses.[1−3]

Large-scale metabolite analytical methods have been developed using gas chromatography−mass spectrometry,[4,5] nuclear magnetic resonance,[6,7] Fourier transform ion cyclotron resonance mass spectrometry,[8] or electrospray ionization mass spectrometry (ESI-MS).[9] We previously reported that by capillary electrophoresis−mass spectrometry (CE−MS) it was possible to detect 1692 peaks from *Bacillus subtilis* extracts.[10] However, due to a lack of commercially available standards, most of the peaks could not be identified. Therefore, in the absence of such standards, novel identification methods are needed.

Several computational approaches have been described to predict the mobility of analytes.[11−24] One such approach uses a semiempirical physicochemical model to estimate electrophoretic

[*] To whom correspondence should be addressed. Tel: (+81) 235 29 0528. Fax: (+81) 235 29 0530. E-mail: soga@sfc.keio.ac.jp.

[†] Keio University.

[‡] Mitsubishi Space Software Co. Ltd.

[§] The University of Tokyo.

[||] National Institute of Advanced Industrial Science and Technology.

[⊥] Human Metabolome Technology Inc.

[#] Kyoto University.

(1) Soga, T.; Heiger, D. N. *Anal. Chem.* **2000**, *72*, 1236−1241.
(2) Soga, T.; Imaizumi, M. *Electrophoresis* **2001**, *22*, 3418−3425.
(3) Soga, T.; Ueno, Y.; Naraoka, H.; Matsuda, K.; Tomita, M.; Nishioka, T. *Anal. Chem.* **2002**, *74*, 6224−6229.
(4) Fiehn, O.; Kopka, J.; Dormann, P.; Altmann, T.; Trethewey, R. N.; Willmitzer, L. *Nat. Biotechnol.* **2000**, *18*, 1157−1161.
(5) Fiehn, O.; Kopka, J.; Trethewey, R. N.; Willmitzer, L. *Anal. Chem.* **2000**, *72*, 3573−3580.
(6) Reo, N. V. *Drug Chem. Toxicol.* **2002**, *25*, 375−382.
(7) Nicholson, J. K.; Wilson, I. D. *Nat. Rev. Drug Discovery* **2003**, *2*, 668−676.
(8) Aharoni, A.; Ric de Vos, C. H.; Verhoeven, H. A.; Maliepaard, C. A.; Kruppa, G.; Bino, R.; Goodenowe, D. B. *Omics* **2002**, *6*, 217−234.
(9) Castrillo, J. I.; Hayes, A.; Mohammed, S.; Gaskell, S. J.; Oliver, S. G. *Phytochemistry* **2003**, *62*, 929−937.
(10) Soga, T.; Ohashi, Y.; Ueno, Y.; Naraoka, H.; Tomita, M.; Nishioka, T. *J Proteome Res.* **2003**, *2*, 488−494.
(11) Fu, S.; Lucy, C. A. *Anal. Chem.* **1998**, *70*, 173−181
(12) Fu, S.; Li, D.; Lucy, C. A. *Analyst* **1998**, *123*, 1487−1492.
(13) Li, D.; Fu, S.; Lucy, C. A. *Anal. Chem.* **1999**, *71*, 687−699.
(14) Li, D.; Lucy, C. A. *Anal. Chem.* **2001**, *73*, 1324−1329.
(15) Jouyban, A.; Chan, H. K.; Clark, B. J.; Kenndler, E. *Electrophoresis* **2003**, *24*, 1596−1602.
(16) Grossman, P. D.; Colburn, J. C.; Lauer, H. H. *Anal. Biochem.* **1989**, *179*, 28−33.
(17) Tessier, B.; Blanchard, F.; Vanderesse, R.; Harscoat, C.; Marc, I. *J. Chromatogr., A* **2004**, *1024*, 255−266.
(18) Yao, X.; Zhang, X.; Zhang, R.; Liu, M.; Hu, Z.; Fan, B. *Talanta* **2002**, *57*, 297−306.
(19) Agatonovic-Kustrin, S.; Zecevic, M.; Zivanovic, L. *J. Pharm. Biomed. Anal.* **1999**, *21*, 95−103.
(20) Tham, S. Y.; Agatonovic-Kustrin, S. *J. Pharm. Biomed. Anal.* **2002**, 28, 581−590.
(21) 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.
(22) Jalali-Heravi, M.; Garkani-Nejad, Z. *J. Chromatogr., A* **2001**, *927*, 211−218.

10.1021/ac048950g CCC: $30.25 © 2005 American Chemical Society
Published on Web 11/26/2004

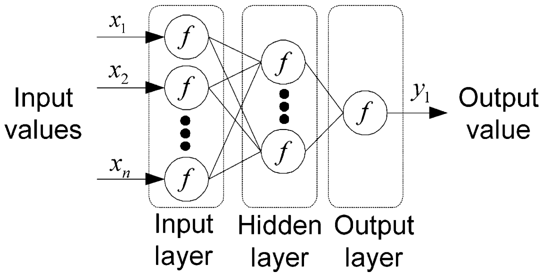

Figure 1. Typical multilayered architecture of ANN. The network is composed of three layers: (1) an input layer, (2) a hidden layer, and (3) an output layer. Each layer is composed of neurons symbolized by circles. Each neuron processes input to output using a sigmoid function. In this study, we employed a single hidden layer and an output layer, making use of only one neuron.

mobility from basic molecular properties such as the molecular size and charge. While the model can be optimized to predict the mobility of a particular group of molecules[11−15] or short peptides,[16,17] its applicability to a more general or varied set of molecules has not been described.

Another prediction method is based on the artificial neural network (ANN).[18−24] This learning algorithm for solving classification or identification problems[25−28] has been used to model the quantitative structure−retention relationships for gas chromatography[18] and the retention time for reversed-phase high-performance liquid chromatography (RP-HPLC)[19−21] or the migration time for capillary electrophoresis.[22−24] Previous predictions[19−24,26,28−32] with ANNs used the multiple-layers architecture model with back-propagation learning (Figure 1); the training methodology encountered the problem of generating many local minimums, which resulted in a decrease in the prediction accuracy. To avoid this problem, we used ANN ensemble techniques[33−35] to predict the electrophoretic mobility of cations in CE−MS. The predicted electrophoretic mobilities of the 10 793 compounds in the KEGG ligand database[36] were validated by comparing them with 241 experimentally measured mobilities.

METHOD

Reagents. 1-Adamantanamine was purchased from ICN Biomedicals (Aurora, OH), L-albizziine from Bachem (Bubendorf, Switzerland), L-cystathionine from Toronto Research Chemicals (Ontario, Canada), and L-methionine sulfone from Avocado (Lancashire, U.K.). All other reagents were obtained from common commercial sources. Individual stock solutions at concentrations of 10 or 100 mM were prepared in Milli-Q water, 0.1 N HCl, or 0.1 N NaOH. The working standard mixture was prepared by diluting these stock solutions with Milli-Q water just prior to injection. All chemicals were of analytical or reagent grade. Water was purified with a Milli-Q purification system (Millipore, Bedford, MA).

Instrumentation. All CE−MS experiments were performed using an Agilent CE capillary electrophoresis system equipped with an air pressure pump, an Agilent 1100 series mass selective detector (MSD) mass spectrometer, an Agilent 1100 series isocratic HPLC pump, a G1603A Agilent CE−MS adapter kit, and a G1607A Agilent CE−ESI-MS sprayer kit (Agilent Technologies, Palo Alto, CA). System control, data acquisition, and MSD data evaluation were performed using G2201AA Agilent ChemStation software for CE-MSD.

CE−MS Conditions. Separations were carried out in a fused-silica capillary (50 μm i.d. × 100 cm total length) using 1 M formic acid as the electrolyte.[1] Samples were pressure-injected at 50 mbar for 3 s (3 nL). The applied voltage was +30 kV. The capillary temperature was maintained at 20 °C using a thermostat, and the sample trays were cooled to below 5 °C. A solution of 5 mM ammonium acetate in 50% (v/v) methanol−water was delivered as the sheath liquid at 10 μL/min. The pH of the solution was 1.8. ESI-MS was conducted in the positive ion mode with the capillary voltage set at 4000 V. The flow of heated dry nitrogen gas (300 °C) was maintained at 10 L/min. MS with selective ion monitoring (SIM) was used to monitor analytes. To minimize migration time variations due to electroosmotic flow fluctuations,[37] the migration times of all species were normalized using L-methionine sulfone as an internal standard.

Brief Description of Artificial Neural Network Ensemble. An ANN consists of fully interconnected layers that include units called neurons. Each unit in the input layer is connected to all other units in the hidden layer, and each unit in the hidden layer is linked to all other units in the output layer. Funahashi[38] previously demonstrated that a single hidden layer could approximate any function. The output of each unit is determined by its input sums through a nonlinear transfer function.

In the back-propagation method[25] to train the weights, the performance of optimization was evaluated by the mean squared error (MSE), which was calculated as

$$\frac{1}{mn}\sum_{k=1}^{n}\sum_{l=1}^{m}(y_{kl} - t_{kl})^2 \qquad (1)$$

where m is the number of training sets, n the number of units in the outputs, y_{kl} the element of the objective output matrix ($m \times n$) for the training set or test set, and t_{kl} the element of the ANN output matrix ($m \times n$). The MSE is propagated through the network from the output to the input layer, and the corresponding weights between units are changed to minimize the error.

[23] Jalali-Heravi, M.; Garkani-Nejad, Z. *J. Chromatogr., A* **2002**, *971*, 207−215.

[24] Malovana, S.; Frias-Garcia, S.; Havel, J. *Electrophoresis* **2002**, *23*, 1815−1821.

[25] Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. *Nature* **1986**, *323*, 533−536.

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

[27] Aires-de-Sousa, J.; Hemmer, M. C.; Gasteiger, J. *Anal. Chem.* **2002**, *74*, 80−90.

[28] Muzikar, M.; Havel, J.; Macka, M. *Electrophoresis* **2003**, *24*, 2252−2258.

[29] Jalali-Heravi, M.; Fatemi, M. H. *J. Chromatogr., A* **1998**, *825*, 161−169.

[30] Jalali-Heravi, M.; Fatemi, M. H. *J. Chromatogr., A* **2000**, *897*, 227−235.

[31] Jalali-Heravi, M.; Parastar, F. *J. Chromatogr., A* **2000**, *903*, 145−154.

[32] Loukas, Y. L. *J. Chromatogr., A* **2000**, *904*, 119−129.

[33] Hansen, L. K.; Salamon, P. *IEEE Trans. Pattern Anal. Mach. Intell.* **1990**, *12*, 993−1001.

[34] Baxt, W. G. *Neural Comput.* **1992**, *4*, 772−780.

[35] Opitz, D. W.; Shavlik, J. W. *Connection Sci.* **1996**, *8*, 337−353.

[36] Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. *Nucleic Acids Res.* **2004**, *32*, D277−280 (Database issue).

[37] Schmitt-Kopplin, P.; Frommberger, M. *Electrophoresis* **2003**, *24*, 3837−3867.

[38] Funahashi, K. *Neural Network* **1989**, *2*, 183−191.

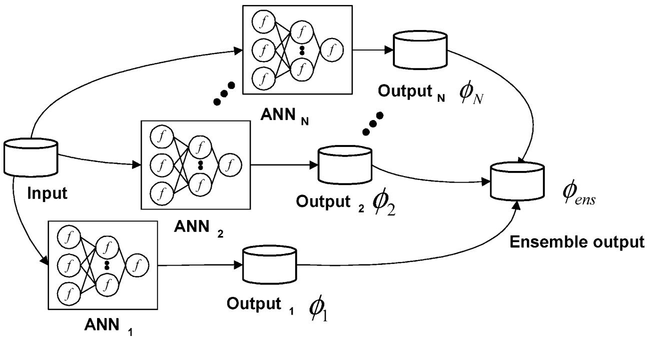

Figure 2. Ensemble method of artificial neural networks. Input data are first used for multiple individual ANNs. The ensemble output is then built by combining the outputs of each ANN.

To improve prediction accuracy, we employed the ANN ensemble method.[33-35] The scheme includes two steps. In the first, the training sets and test sets are assigned for several ANNs, and the ANNs are trained with different initial weights. In the second step, the outputs of the ANNs are collected and averaged as the final predictor:

$$\phi_{\rm ens}(x) = \frac{1}{N}\sum_{i=1}^{N}\phi_i(x) \qquad (2)$$

where $\phi_{\rm ens}(x)$ is the ensembled output, N the number of ANNs to be combined, and $\phi_i(x)$ the output of ANN. The scheme of the ANN ensemble is shown in Figure 2.

Bagging[39] and boosting[40,41] methods are used to generate and assemble multiple ANNs. Briefly, bagging randomly selects training sets to create different training inputs. Boosting generates a sequence of ANNs whose training sets are selected according to the performance of the preceding ANNs. Since the performance of bagging and boosting may depend on the particular problem considered,[42,43] we employed a method to average predictions of several ANNs derived from the original ANN ensemble principle.[33] This versatile method was reported as simple but sufficiently accurate. [42]

Inputs of Artificial Neural Networks. The ANN training set consisted of a net charge under the measurement condition (pH 1.8), pK_a value, and 152 molecular descriptors (see below for details). Using the Henderson−Hasselbach equation, the positive charge α_b is computed as in eq 3 and the negative charge α_a, as in eq 4:

$$\alpha_{\rm b} = \sum_{i}^{n}\frac{1}{1+10^{({\rm p}Ka i-{\rm pH})}} \qquad (3)$$

$$\alpha_{\rm a} = \sum_{j}^{m}\frac{1}{1+10^{({\rm pH}-{\rm p}Ka j)}} \qquad (4)$$

The net charge of zwitterions or ampholytes q is calculated using the equation

$$q = \alpha_{\rm a} - \alpha_{\rm b} \qquad (5)$$

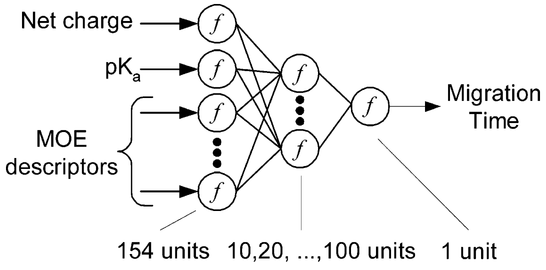

Figure 3. Specific ANN architecture used in the present study. The net charge on the molecule, the pK_a value closest to 1.8 (when more than one pK_a exists), and multiple MOE descriptors were employed as ANN inputs. The ANN output was the RMT. All inputs and the output were assigned normalized values.

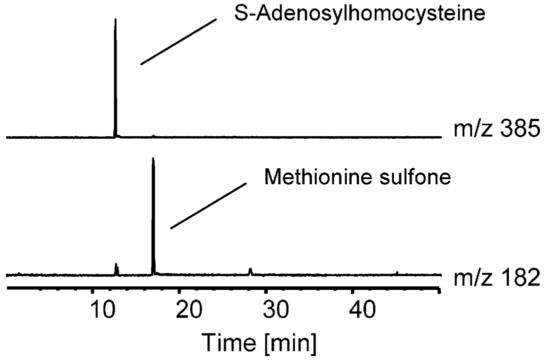

Figure 4. Selected ion electropherograms of *S*-adenosylhomocysteine and the internal standard methionine sulfone. These show a typical extracted electropherogram obtained by CE−MS in the SIM mode. The migration time of the analyte is normalized by dividing its value by that of the internal standard, methionine sulfate, to give the RMT, which is independent of the variability in separation conditions between various consecutive sample runs.

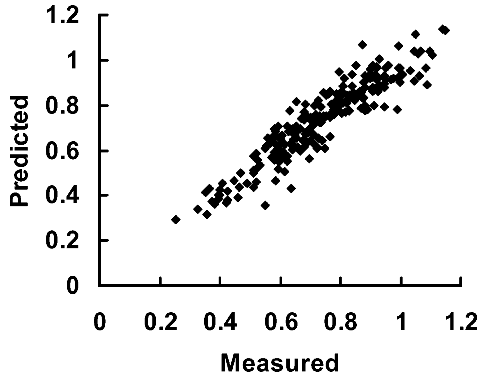

Figure 5. Correlation between experimentally measured and relative migration times of all analyzed cations predicted by the ANN ensemble.

where i and j are suffixes of pK_a, n is the number of pK_a values of a negatively charged species, and m is the number of pK_a values of a positively charged species.[44] The pK_a values were obtained from ACD/Lab pK_a DB.[45] The 152 molecular descriptors were

(39) Breinman, L. *Machine Learn.* **1996**, *24*, 123−140.

(40) Schapire, R. E. *Machine Learn.* **1990**, *5*, 197−227.

(41) Freund, Y. *Inf. Comput.* **1995**, *121*, 256−285.

(42) Opitz, D. W.; Maclin, R. *J. Artif. Intelligence Res.* **1999**, *11*, 169−198.

(43) Agrafiotis, D. K.; Cedeno, W.; Lobanov, V. S. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903−911.

(44) Skoog, B.; Wichman, A. *Trends Anal. Chem.* **1986**, *5*, 82−83.

Table 1. Experimentally Measured and Predicted RMTs of All 241 Selected Cations

name	RMT		rank	no. of compds[a]	MW	name	RMT		rank	no. of compds[a]	MW
	meas	pred					meas	pred			
isopropylamine	0.599	0.550	2	15	59	1-aminocyclopentane-carboxylate	0.460	0.392	1	23	129
pyrazole	0.617	0.505	2	3	68	metformin	0.673	0.742	4	23	129
α-aminopropiononitrile	0.512	0.436	1	6	70	agmatine	0.652	0.655	1	36	130
methylguanidine	0.525	0.524	1	13	73	*N*-acetylputrescine	0.428	0.420	1	36	130
1,3-diaminopropane	0.375	0.375	1	19	74	L-leucine	0.688	0.684	1	37	131
glycine	0.702	0.682	1	13	75	creatine	1.006	0.936	6	37	131
trimethylamine *N*-oxide	0.586	0.575	1	13	75	L-isoleucine	0.920	0.962	6	37	131
semicarbazide	0.582	0.632	4	13	75	5-aminolevulinate	0.747	0.608	14	37	131
cysteamine	0.550	0.358	1	3	77	*trans*-4-hydroxy-L-proline	0.575	0.663	2	37	131
pyridine	0.512	0.509	1	2	79	L-norleucine	0.867	0.838	5	37	131
3-amino-1,2,4-triazole	0.520	0.584	1	8	84	6-aminohexanoate	0.858	0.854	2	37	131
piperidine	0.550	0.608	1	3	85	3-guanidinopropanoate	0.869	0.855	1	37	131
piperazine	0.367	0.429	1	16	86	bis(3-aminopropyl)amine	0.653	0.705	4	37	131
3-methylbutanamine	0.634	0.777	2	8	87	*cis*−4-hydroxy-L-proline	0.357	0.314	1	37	131
putrescine	0.399	0.403	1	15	88	L-ornithine	0.881	0.879	1	37	132
L-alanine	0.762	0.851	4	14	89	L-asparagine	0.682	0.805	8	37	132
β-alanine	0.619	0.566	2	14	89	glycylglycine	0.577	0.563	1	37	132
sarcosine	0.787	0.806	1	14	89	4-aminoindole	0.642	0.695	5	37	132
3-aminopropane-1,2-diol	0.594	0.517	1	5	91	5-aminoindole	0.627	0.682	4	37	132
aniline	0.639	0.604	1	5	93	L-aspartate	0.969	0.915	2	17	133
1-aminocyclopropane-1-carboxylate	0.736	0.721	1	14	101	2-aminobenzimidazole	0.607	0.592	1	17	133
hexylamine	0.657	0.615	2	14	101	adenine	0.643	0.651	1	11	135
cadaverine	0.407	0.453	1	17	102	L-homocysteine	0.850	0.979	3	11	135
4-aminobutanoate	0.649	0.664	1	21	103	hypoxanthine	0.913	0.797	1	40	136
N,*N*-dimethylglycine	0.927	0.859	2	21	103	anthranilate	0.890	0.919	2	9	137
N-methyl-L-alanine	0.877	0.899	3	21	103	tyramine	0.805	0.863	1	9	137
2-amino-2-methyl-propanoate	0.797	0.773	1	21	103	4-aminobenzoate	0.609	0.573	2	9	137
2,3-diaminopropanoate	0.594	0.600	1	18	104	phenylethanolamine	0.524	0.555	1	9	137
L-serine	0.840	0.836	1	7	105	1-methylnicotinamide	0.698	0.693	1	9	137
diethanolamine	0.624	0.551	3	7	105	pralidoxime	0.714	0.718	1	9	137
n-methylaniline	0.606	0.663	2	7	107	urocanate	0.708	0.673	2	14	138
phenylhydrazine	0.553	0.621	1	12	108	tropinone	0.880	0.777	2	13	139
aminophenol	0.666	0.692	2	9	109	*N*-methylnicotinate	0.665	0.595	3	13	139
2-aminophenol	0.651	0.703	4	9	109	tropine	0.568	0.632	6	18	141
4-hydroxyaniline	0.690	0.676	4	9	109	5-(2-hydroxyethyl)-4-methylthiazole	0.688	0.634	2	7	143
isoguvacine (RBI)	0.713	0.797	7	9	109	ectoine	0.788	0.788	1	12	144
cytosine	0.599	0.657	2	15	111	spermidine	0.726	0.655	6	29	145
1*H*-imidazole-4-ethanamine	0.401	0.425	1	15	111	*γ*-guanidinobutyrate	0.384	0.359	1	29	145
creatinine	0.595	0.554	1	12	113	L-lysine	0.877	0.831	3	38	146
L-proline	0.907	0.863	1	4	115	D-glutamine	0.582	0.593	2	38	146
L-valine	0.639	0.680	2	18	117	dimethylbenzimidazole	0.658	0.694	1	38	146
5-aminopentanoate	0.842	0.856	4	18	117	trientine	0.255	0.293	1	38	146
guanidinoacetate	0.682	0.647	2	18	117	L-glutamate	0.918	0.938	6	21	147
L-norvaline	0.858	0.844	4	18	117	*O*-acetyl-L-serine	1.060	0.933	4	21	147
benzimidazole	0.588	0.617	3	13	118	L-albizziine	0.919	0.913	2	21	147
l-2,4-diaminobutanoate	0.592	0.577	2	13	118	L-methionine	0.638	0.630	2	34	149
L-threonine	0.898	0.916	2	15	119	3-methyladenine	0.657	0.608	6	34	149
L-homoserine	0.847	0.823	3	15	119	1-methyladenine	0.587	0.681	5	34	149
2-methylserine	0.872	0.833	3	15	119	triethanolamine	0.900	0.937	4	34	149
L-allothreonine	0.906	0.861	3	15	119	penicillamine	0.996	0.900	5	34	149
purine	0.716	0.740	1	26	120	guanine	0.700	0.744	1	20	151
benzamidine	0.584	0.664	1	26	120	1-adamantanamine	0.722	0.613	7	20	151
L-cysteine	0.947	0.879	2	13	121	3-hydroxyanthranilate	0.901	0.876	3	16	153
1-phenylethylamine	0.675	0.643	3	13	121	4-aminosalicylate	1.089	0.892	4	16	153
N,*N*-dimethylaniline	0.612	0.589	2	13	121	4-(2-aminoethyl)−1,2-benzenediol	0.893	0.931	4	16	153
phenethylamine	0.681	0.613	4	13	121	mesalamine	0.733	0.723	1	16	153
2,6-dimethylaniline	0.739	0.645	2	13	121	pseudopelletierine	0.702	0.703	1	16	153
nicotinamide	0.595	0.708	2	15	122	*Nω*-acetylhistamine	0.659	0.698	2	16	153
picolinamide	0.882	0.798	1	15	122	L-histidine	0.617	0.626	2	15	155
3-chloro-L-alanine	1.082	0.967	2	17	124	buformin	0.511	0.500	1	10	157
5-methylcytosine	0.686	0.674	1	10	125	*N*-methyl-*N*-propagyl-benzylamine	0.716	0.624	4	17	159
n-methylhistamine	0.395	0.382	1	10	125	tryptamine	0.700	0.732	2	31	160
imidazole-4-acetate	0.655	0.813	2	16	126	tolazoline	0.582	0.697	3	31	160
L-pipecolate	0.866	0.860	1	23	129	l-2-aminoadipate	0.907	0.922	2	18	161
octylamine	0.794	0.840	1	23	129	*N*-methyl-L-glutamate	1.059	1.028	1	18	161

Table 1 (Continued)

	RMT			no. of	
name	meas	pred	rank	compds[a]	MW
O-acetyl-L-homoserine	0.921	0.970	4	18	161
L-carnitine	0.599	0.618	1	53	162
nicotine	0.700	0.641	2	53	162
procollagen 5-hydroxy-L-lysine	0.351	0.411	1	53	162
anabasine	0.469	0.499	1	53	162
pterin	0.797	0.947	5	21	163
L-phenylalanine	0.698	0.565	8	21	165
3-methylguanine	0.685	0.755	2	21	165
7-methylguanine	0.929	0.889	1	21	165
benzocaine	0.823	0.846	4	21	165
ethionamide	0.717	0.758	1	34	166
pyridoxal	0.695	0.688	1	21	167
phenylephrine	0.991	0.783	8	21	167
epinine	0.752	0.845	10	21	167
mecamylamine	0.737	0.734	2	21	167
pyridoxamine	0.523	0.458	1	39	168
pyridoxine	0.626	0.708	6	18	169
N(π)-methyl-L-histidine	0.624	0.573	3	18	169
3-methyl-L-histidine	0.638	0.625	2	18	169
metronidazole	0.737	0.794	2	17	171
1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine	0.680	0.646	3	18	173
L-arginine	0.788	0.823	2	30	174
*N*2−acetyl-L-ornithine	0.948	0.897	1	30	174
*N*5-ethyl-L-glutamine	0.602	0.647	1	30	174
gramine	0.708	0.727	2	30	174
L-citrulline	0.926	0.869	2	21	175
3-(2-aminoethyl)−1*H*-indol-5-ol	0.752	0.750	1	35	176
D-glucosamine	0.745	0.853	9	40	179
D-galactosamine	0.766	0.859	10	40	179
1,10-phenanthroline	0.561	0.656	2	64	180
L-tyrosine	0.950	0.914	2	14	181
dacarbazine	0.739	0.733	1	40	182
harman	0.698	0.693	1	40	182
N_8-acetylspermidine	0.636	0.431	5	15	187
amrinone	0.532	0.536	2	15	187
castanospermine	0.769	0.764	1	20	189
ll-2,6-diaminoheptanedioate	0.730	0.751	2	28	190
L-tyrosine methyl ester	0.998	0.967	2	14	195
2-amino-2-deoxy-D-gluconate	0.778	0.779	1	14	195
3,4-dihydroxy-L-phenylalanine	0.976	0.919	1	20	197
N-acetyl-L-histidine	0.822	0.810	1	20	197
tacrine	0.765	0.658	3	24	198
thiabendazole	0.654	0.684	1	11	201
spermine	0.512	0.577	2	31	202
arginine ethyl ester	0.329	0.340	1	31	202
L-tryptophan	0.923	0.919	3	50	204
O-acetylcarnitine	0.727	0.821	7	50	204
L-kynurenine	0.813	0.919	3	20	208
minoxidil	0.785	0.767	1	5	209
zalcitabine	0.681	0.734	2	19	211
milrinone	0.745	0.853	5	19	211
harmaline	0.746	0.753	3	24	214
kinetin	0.766	0.747	1	6	215
6-aminopenicillanate	0.900	0.961	2	19	216
β-alanyl-L-lysine	0.573	0.640	1	10	217
zeatin	0.804	0.764	4	15	219
5-hydroxy-L-tryptophan	0.946	0.933	1	31	220
dihydrozeatin	0.810	0.810	1	24	221
L-cystathionine	0.819	0.770	2	34	222
3-hydroxy-L-kynurenine	0.800	0.839	2	23	224
acyclovir	0.947	0.791	5	12	225
6-benzylaminopurine	0.786	0.782	1	12	225
carnosine	0.571	0.570	1	20	226
fenfluramine	0.724	0.608	2	14	231
aminopyrine	0.789	0.848	2	14	231
homocarnosine	0.576	0.576	2	26	240
cytidine	0.777	0.759	3	14	243
N_1-acetylspermine	0.462	0.435	1	41	244
N_2-(D,L-carboxyethyl)-L-arginine	0.908	0.879	1	38	246
pyridoxamine phosphate	0.883	0.932	1	39	248
deoxyadenosine	0.797	0.796	1	12	251
5′-deoxyadenosine	0.807	0.883	3	12	251
nebularine	1.047	0.904	1	16	252
L-arginine phosphate	1.066	1.038	2	33	254
thiamine	0.847	0.848	1	11	265
albendazole	0.581	0.559	1	11	265
adenosine	0.829	0.843	2	21	267
deoxyguanosine	0.960	0.975	3	21	267
physostigmine	0.825	0.840	2	20	275
Glu−Glu	0.875	1.070	7	28	276
1-methyladenosine	0.770	0.802	2	16	281
guanosine	1.045	1.041	1	17	283
atropine	0.825	0.778	3	16	289
L-hyoscyamine	0.831	0.778	4	16	289
trimethoprim	0.837	0.882	3	43	290
5′-methylthioadenosine	0.821	0.819	1	16	297
scopolamine	0.583	0.468	1	16	303
anserine nitrate	0.838	0.936	5	16	303
D-nopaline	0.995	0.916	5	24	304
glutathione	1.103	1.024	3	13	307
bestatin	0.928	1.007	4	22	308
neamine	0.421	0.367	1	20	322
piroxicam	1.098	1.038	2	18	331
thiamine monophosphate	0.908	0.974	2	14	345
protopine	0.868	0.784	4	13	353
tetrahydropalmatine	0.910	0.931	4	10	355
allocryptopine	0.891	0.781	4	12	369
(*R*)-*S*-lactoylglutathione	1.146	1.133	1	12	379
S-adenosyl-L-homocysteine	0.679	0.769	2	15	384
Leu-Leu-Tyr	0.996	0.962	4	21	408
leupeptin	0.994	1.063	1	24	427
3,5-diiodo-L-tyrosine	1.051	1.113	1	13	433
vistamycin	0.996	0.922	1	8	454
methotrexate	0.492	0.452	1	8	454
psychosine	1.031	0.954	4	13	462
tobramycin	0.395	0.401	1	16	468
puromycin	0.712	0.711	1	4	472
kanamycin B	0.425	0.381	1	12	484
apramycin	0.445	0.463	1	4	540
adriamycin	0.952	0.964	1	8	544
triiodothyronine	1.138	1.138	1	7	651

[a] Represents the number of compounds with molecular weights in the corresponding molecular weight (MW) column as listed in KEGG.

calculated by the Molecular Operating Environment (MOE).[46] First, we estimated a 3D molecular structure from the MOL format (MDL Information System, San Leandro, CA) for each metabolite using the energy minimization function and calculated 192

[45] ACD/pK_a DB, version 7.07, Advanced Chemistry Development, Inc., Toronto ON, Canada, www.acdlabs.com, 2003.

[46] Chemical Computing Group, Molecular Operating Environment (MOE) 2003.02; 1225 University Street, Montreal, PQ, Canada.

molecular descriptors provided by MOE. Among these, 40 were eliminated since they produced duplicate values for our 241 target compounds. The remaining 152 descriptors listed in Table S1 (Supporting Information) were used as ANN inputs. A detailed description of the descriptors is available in the MOE documentation. Then, the ANN was trained using the descriptors, the net charge, and the pK_a value as inputs and the relative migration times (RMTs) as outputs. The output values were normalized by dividing retention time values by the migration time of methionine sulfone.

ANN Learning. A cross-validation method was used to evaluate the prediction performance of the ANN ensemble. A total of 241 cationic metabolites were randomly divided into the training set (90% of the metabolites) and the test set (the remaining 10%). The evaluation procedure was repeated 10 times until all metabolites had been selected at least once as a member of the test sets. The values applied for inputs or outputs of ANN were normalized to a numerical value between 0.1 and 0.9. For the nonnumerical descriptors such as infinity calculated by MOE, substituted 0 was used for ANN inputs.

The following parameters were used for the ANN: learning rate with values of 0.02, 0.03, ..., 0.07 used to control convergence speed; different numbers of hidden units ranging 10, 20, ..., 100; the learning momentum, used to prevent divergence during the training phase, was set at 0.9; the terminal epoch, which is the maximum number of times to update weights, was fixed at 20 000. Weights were initialized using random values between −0.3 and 0.3. The overall architecture of the ANN used in our study is depicted in Figure 3. After 10 individual ANNs were trained using different initial weights, an ANN ensemble was built by combining the 30 trained ANN predictors with the highest correlation coefficients.

RESULTS AND DISCUSSION

Prediction of Cation RMTs. Generally, the parameters for training ANNs (e.g., the training ratio or the number of hidden units) are determined empirically. Because the gain in prediction accuracy of an ANN ensemble depends on the disagreement among the predicted values by the assembled ANNs,[33,35,43] to raise accuracy, it is preferable to train ANNs on different conditions. Therefore, we trained the ANN with different learning parameters.

It is also desirable for ANN inputs to consider all features that may influence molecular mobility. While all molecules have the same number of descriptors and one net charge value, their number of different pK_a values varies. Most molecules have only one or two pK_a values; however, a few have more than 10 pK_as. In the present work, we selected only one pK_a value that is most relevant to the experimental conditions (pH 1.8) to allow for the fixed number of ANN inputs.

Figure 4 shows CE−MS electropherograms of *S*-adenosylhomocysteine as obtained with the single-ion monitoring mode at m/z 385 and of methionine sulfone (internal standard) at m/z 182. The RMT of *S*-adenosylhomocysteine was calculated using the RMT of methionine sulfone. Under the conditions used, cations were predominantly detected as their protonated $[M + H]^+$ molecular ions.[1] The RMT of 241 cations was measured and calculated in the same way.

Table 2. Correlation Coefficient Calculated Using the Variable Descriptor Sets[a]

MOE descriptors	all				2D with VSA	only VSA
net charge	+	+	−	−	+	+
pK_a	+	−	+	−	+	+
correlation coefficient	0.931	0.929	0.855	0.855	0.927	0.921

[a] “+” and “−“ reflect the use or nonuse, respectively, of the molecular descriptor.

Figure 5 is a comparison between predicted and measured RMTs for 241 cations (also see Table 1). The correlation coefficients of the results of ANN ensemble and single ANN were 0.931 and 0.905, respectively. The result predicted by the ANN ensemble showed a better correlation than the prediction by a single ANN. Most RMT values (178 of 241) were found within the 0.6−1.0 range, forming a normal distribution (mean 0.73; variance 0.18). The mean prediction error was 0.049. To reveal important factors in the prediction, the ANN ensemble was tested with a reduced set of inputs (Table 2). While removal of the pK_a value had little effect on the correlation coefficient, removal of the net charge value dramatically decreased the coefficient. The lower significance of the pK_a value probably results from our manual selection of a single pK_a value for each compound to accommodate for the fixed number of input nodes in ANN. Although the ANN could not extract the relation between migration time and pK_a values in our experiment, the value calculated from all pK_a values may be more effective in predicting migration time. Among the molecular descriptors, learning with only van der Waals surface area (VSA) descriptors that are calculated from connection table information alone, assuming a dictionary of van der Waals radii and ideal bond lengths,[47] produced a high correlation coefficient of 0.921. The addition of 2D and 3D descriptors increased the coefficient by only 0.006 and 0.010, respectively. Thus, we considered VSA descriptors most effective for the prediction of migration times.

Identification of Cations. The 241 experimentally measured RMTs were compared with the predicted RMT values of 10 793 molecules in the KEGG ligand database. The rank associated with each genuine molecule is plotted in Figure 6. Among 241 cations, 102 (42.3%) had the RMT values closest to those of the correct molecules; 188 (78.0%) were within the 3 closest RMT values among 10 793 candidates. An investigation of mispredictions revealed that the RMTs of optical isomers were differently predicted by the present method. For instance, the RMTs of L-isoleucine and D-isoleucine were predicted as 0.854 and 0.823, respectively, although all optical isomers are experimentally indistinguishable (also see Table S2 in the Supporting Information). This prediction error originated in the 3D descriptors computed through the energy minimization function of MOE. For large compounds, the associated increase in error will have a greater influence on the prediction of RMTs. To investigate the effect of errors in the prediction, we first measured RMTs of 102

(47) Labute, P. *J. Mol. Graphics Modell.* **2000**, *18*, 464−477.

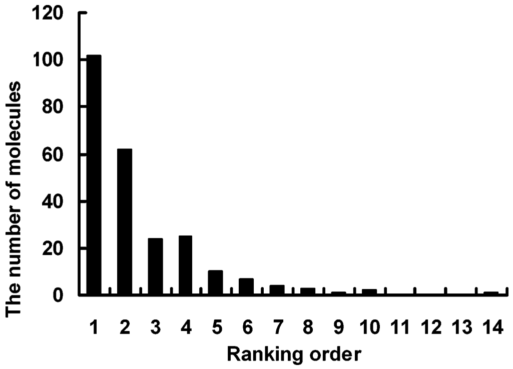

Figure 6. Ranking of analyte identification success using RMT estimates. Data from the 241 test compounds are shown. The ranking number corresponds to the position of a compound in the list of candidates when all possible molecules of a given *m*/*z* value are sorted in order of increasing difference between experimental and predicted RMT. For example, for a given experimental RMT, a ranking order of 2 will result if the analyte under study corresponds to the compound having the second smallest difference between the experimental and predicted RMT at the *m*/*z* value of the unknown analyte. The number of individual compounds found in each rank is shown.

Table 3. Comparison between the Prediction Accuracy of the ANN Ensemble and a Single ANN[a]

	measured data		noised data	
ANN	ensemble	single	ensemble	single
correlation coefficient	0.931	0.905	0.915	0.892
mean prediction error (RMT)	0.049	0.057	0.055	0.063
mean ID ranking				
integer[a]	2.427	2.548	2.595	2.727
first decimal place[b]	1.975	2.071	2.214	2.159
second decimal place[b]	1.676	1.722	1.773	1.784
third decimal place[b]	1.560	1.573	1.603	1.609

[a] Noised data were prepared artificially using random numbers. The average of identification prediction ranking was calculated from the candidate molecules, considering different degrees of precision of their weights. The results calculated actual measured data listed in the "measured data" column. The results in the "noised data" column were calculated by the artificially noised RMTs. [b] Molecular weight precision.

randomly selected molecules more than 4 times. The errors of RMTs distributed normally; the deviation was 0.019. To assess the prediction accuracy in a noisy environment, we prepared artificial RMT data doped with normally distributed noises. The prediction with the noisy data was conducted 5 times using different noises. Table 3 summarizes the average results. Although the noise slightly decreased the overall prediction accuracy in both ANN ensemble and single ANN, the superiority of the ANN ensemble was unchanged. Since the prediction accuracy did not deteriorate much as a result of the artificial errors, we concluded that our prediction results are robust to noise.

Generally, the performance of the single ANN is susceptible to the order of learning data and the initial weights, but the performance of the ANN ensemble is resistant to such factors and noises. Therefore, the ANN ensemble is better suited for improving identification capability.

CONCLUSIONS

A computational method based on an ANN ensemble was developed for the prediction of electrophoretic mobility in CE–MS. For a total of 241 commercially available cations, RMTs were experimentally measured by CE–MS and also estimated using a set of 154 calculated molecular descriptors. The results demonstrate that the estimation method based purely on computational analysis made possible the prediction of the RT of cations with a very good correlation coefficient. The utility of our method was demonstrated by predicting RMTs of all metabolites listed in the KEGG ligand database and by matching the results with experimentally determined RMTs of 241 different cationic metabolites. In addition, when used for unknown peak characterization, for a given *m*/*z* and experimental RMT, the correct compound was found within the top three candidates in 78.0% of cases. These results indicate that this approach can be used for the prediction of the migration time of any cation and that it is a powerful new tool to help identify the compounds present in uncharacterized CE or LC peaks during metabolome analysis.

ACKNOWLEDGMENT

We thank all members of our laboratory (Y. Kakazu, Y. Ueno, K. Sugawara, S. Igarashi, and S. Ikeda) for experimental data collection and technical advice. We also thank Dr. Martin Robert, IAB, Keio University, for critical reading of the manuscript. This work was supported by a grant from the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-aid for the 21the century Center of Excellence (COE) program entitled 'Understanding and control of life's function via systems biology' (Keio University), a grant from New Energy and Industrial Technology Development and Organization (NEDO) of the Ministry of Economy, Trade and Industry of Japan (Development of a Technological Infrastructure for Industrial Bioprocess Project), a grant from Leading Project for Biosimulation, Ministry of Education, Culture, Sports, Science and Technology, and a grant from Japan Science and Technology Agency (JST).

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 July 18, 2004. Accepted October 8, 2004.

AC048950G