where

$$K'_{s,x} = K_{s,x}\left(1 + \sum_{k_x}\left(\frac{K_{k_x,x}}{c_{k_x}}\right)^{n_{k_x,x}} + \sum_{b}\left(\frac{c_{b_x}}{K_{b_x,x}}\right)^{n_{b_x,x}}\right)$$

$K_{s,x}$ and $n_{s,x}$ are the binding constant and the *cooperativity index* (essentially a Hill exponent) of substance (or effector) s of enzyme x (equilibrium constant for dissociation of the enzyme-ligand complex), respectively, and $K'_{s,x}$ is the former's effective binding constant, which reflects the activities of the competitive activators k_x and the competitive inhibitors b_x; c_s is the concentration of substance s; and k'_x and b'_x are the noncompetitive activators and noncompetitive inhibitors of the reaction catalyzed by enzyme x, respectively [27].

A1.1.10 System N

Glutamine is transported into the cytoplasm by a sodium-dependent transport mechanism. This process is inhibited by histidine [32]:

$$v_{SysN} = V_{max,SysN}\left[\left(\frac{[Na^+]_e}{[Na^+]_e + K_{mNa,SysN}}\right)\left(\frac{[Glu]_e}{[Glu]_e + K_{mGlu,SysN}\left(1 + \frac{[His]_e}{K_{iHis,SysN}}\right)}\right) - \left(\frac{[Na^+]_c}{[Na^+]_c + K_{mNa,SysN}}\right)\left(\frac{[Glu]_c}{[Glu]_c + K_{mGlu,SysN}\left(1 + \frac{[His]_c}{K_{iHis,SysN}}\right)}\right)\right]$$

A1.1.11 System L

Glutamine is transported into the cytoplasm by a sodium-independent transport mechanism. This process is inhibited by tryptophan [32].

$$v_{SysL} = V_{max,SysL}\left(\frac{[Glu]_e}{[Glu]_e + K_{mGlu,SysL}\left(1 + \frac{[Trp]_e}{K_{iTrp,SysL}}\right)} - \frac{[Glu]_c}{[Glu]_c + K_{mGlu,SysL}\left(1 + \frac{[Trp]_c}{K_{iTrp,SysL}}\right)}\right)$$

A1.1.12 Ammonia Transport between Sinusoid and Cytoplasm

Ammonia transport between the sinusoid and cytoplasm was modeled based on the general mass action law:

$$v_{NH_4^+4\text{-}tp} = k_{NH_4^+4\text{-}tp}\left([NH_4^+]_c - [NH_4^+]_e\right)$$

A1.1.13 Transportation of Glutamine, Arginine, and Ammonia between Cytoplasm and Mitochondria

Transports of glutamine, arginine, and ammonia across the mitochondrial membrane were presumed to rapidly attain equilibrium:

$$K_{eq,x}([S]_c - v_x) = ([S]_m + v_x)$$

A1.1.14 Urea Transport to Sinusoid

Excretion of urea in the sinusoidal space was modeled based on the general mass action law:

$$v_{\text{Urea-tp}} = k_{\text{Urea-tp}}\left([\text{urea}]_c - [\text{urea}]_e\right)$$

A1.1.15 Glutamate Transport between Sinusoid and Cytoplasm

Glutamate transport between the sinusoid and cytoplasm was modeled as Michaelis-Menten reversible kinetics:

$$v_{\text{Glu-tp}} = V_{mF,\text{Glu-tp}}\left(\frac{[\text{Glu}]_e}{[\text{Glu}]_e + K_{m\text{Glu,Glu-tp}}}\right) - V_{mR,\text{Glu-tp}}\left(\frac{[\text{Glu}]_c}{[\text{Glu}]_c + K_{m\text{Glu,Glu-tp}}}\right)$$

A1.1.16 Glutamate Flux from the Outside Pathways

Glutamate flux from the outside pathways of the model was represented by the difference between zero-order influx and efflux based on the general mass action law:

$$v_{\text{Glu-spp}} = J_{\text{Glu-spp}} - k_{\text{Glu-spp}}[\text{Glu}]_c$$

A1.1.17 Degradation of Metabolites

Degradation of *N*-acetyl glutamate, Pi, and CoA was modeled based on the general mass action law under the assumption of steady state:

$$v_{\text{deg},s} = k_{\text{deg},s}[s]$$

where s is a substance.

A1.1.18 Ornithine Inflow from Other Reactions

To hold the steady state, ornithine inflow from other reactions was presumed to be equal to the flux of ornithine aminotransferase, v_{OAT}.

A1.2 Mathematical Model of Metabolite Flows in Sinusoid

Flows of ammonia, glutamine, glutamate, and urea from the nth sinusoidal compartment to the n+1th compartment, v_{e,s_n}, were modeled based on the general mass action law:

$$v_{e,s_n} = k_e[s_n]_e$$

where s_n represents a substance in the nth compartment of the sinusoid.

A1.3 Mathematical Model of Gene Expression of Carbamoyl Phosphate Synthetase, Glutamine Synthetase, and Ornithine Aminotransferase in Hepatic Lobule

To describe the regulated gene expression of three enzymes—carbamoyl phosphate synthetase, glutamine synthetase, and ornithine aminotransferase—along the porto-central axis, we adopted the

mechanistic model proposed by Christoffels et al. [5]. The model is based on simple receptor-ligand kinetics, and the parameters are fitted by experimental values. $[F_x^*]$ is the concentration of the active transcription factor F of enzyme x, and assumed as follows [5]:

Carbamoyl phosphate synthetase: $[F_{CPS}^*] = 0.2 - 0.01X$

Glutamine synthetase and ornithine aminotransferase: $[F_{GS}^*] = [F_{OAT}^*] = 0.1X$

where X is the radius of the hepatic lobule: $X = 0$ corresponds to the portal tracts, and $X = 10$ corresponds to the central vein. Thus, X was defined as follows in our model:

$$X = 10 \times \frac{n}{\text{total number of sinusoidal compartments}}$$

where n is the number of a compartment among the eight compartments, $n = 1$ corresponds to the compartment adjacent to the portal tracts, and $n = 8$ corresponds to the compartment adjacent to the central vein. The total number of sinusoidal compartments is eight in our model.

$R_{GX,x}$ is the relative rate of transcription, assumed to correspond to the transcription rate in our model. $R_{GX,x}$ is calculated using the fractional saturation $Y_{GX,x}$, the dissociation constant $K_{GX,x}$, and the Hill coefficient $n_{GX,x}$ as follows [5]:

$$Y_{GX,x} = \frac{[F_x^*]^{n_{GX,x}}}{[F_x^*]^{n_{GX,x}} + K_{GX,x}^{n_{GX,x}}}$$

$$R_{GX,x} = R_{max,GX,x} Y_{GX,x}$$

Carbamoyl phosphate synthetase was fitted with high-affinity ($Y_{GX,CPS,h}$) and low-affinity ($Y_{GX,CPS,l}$) units as follow [5]:

$$R_{GX,CPS} = R_{max,GX,CPS}\left(Y_{GX,CPS,h} + Y_{GX,CPS,l}\right)$$

A1.4 Varying the Uncertain Parameters

The rate constants for glutamate supply from other pathways (the glutamate transport system and the sinusoidal flow model) were uncertain. Therefore we prepared 60 model instances for each type by varying these rate constant values under a steady-state assumption.

Figures 4 and 5 presented the results under the conditions in Table 3 as a representative of the 60 model instances in each gene expression pattern; after 50,000 s from the start of simulation, with the value 3E−5 M s^{-1} for the glutamate influx from pathways outside of the model, the ratio of $V_{mF,Glu\text{-}tp}$ and $V_{mR,Glu\text{-}tp}$ were set to 4.15 in the glutamate transport system, and $k_v = 1.0$ in the sinusoidal flow model.

Appendix 2: Abbreviations

CPS, carbamoyl phosphate synthetase; GS, glutamine synthetase; OAT, ornithine aminotransferase; AGS, *N*-acetyl glutamate synthetase; Glnase, phosphate-dependent glutaminase; OCT, ornithine carbamoyltransferase; ASS, argininosuccinate synthetase; ASL, argininosuccinate lyase; Argase, arginase;

Table 3. Variation parameters.

Regulation of gene expression
1. Not incorporated (N model)
2. Incorporated GS, CPS, and OAT gradients (GCO model)
3. Incorporated only GS gradients (G model)
4. Incorporated GS and CPS gradients (GC model)
5. Incorporated OAT gradients (O model)
6. Incorporated GS and OAT gradients (GO model)
Glutamate transporter
1. $V_{mF,Glu\text{-}tp}$: $V_{mR,Glu\text{-}tp} = 4.15$ ($V_{mF,Glu\text{-}tp} = 1.0629E{-}2\ M\ s^{-1}$, $V_{mR,Glu\text{-}tp} = 2.5611E{-}3\ M\ s^{-1}$)
2. $V_{mF,Glu\text{-}tp}$: $V_{mR,Glu\text{-}tp} = 4.5$ ($V_{mF,Glu\text{-}tp} = 1.2573E{-}3\ M\ s^{-1}$, $V_{mR,Glu\text{-}tp} = 2.7940E{-}4\ M\ s^{-1}$)
3. $V_{mF,Glu\text{-}tp}$: $V_{mR,Glu\text{-}tp} = 5.0$ ($V_{mF,Glu\text{-}tp} = 6.1467E{-}4\ M\ s^{-1}$, $V_{mR,Glu\text{-}tp} = 1.2293E{-}4\ M\ s^{-1}$)
4. $V_{mF,Glu\text{-}tp}$: $V_{mR,Glu\text{-}tp} = 7.0$ ($V_{mF,Glu\text{-}tp} = 2.6560E{-}4\ M\ s^{-1}$, $V_{mR,Glu\text{-}tp} = 3.7943E{-}5\ M\ s^{-1}$)
Glutamate Flux from Outside Pathways
1. $J_{Glu\text{-}spp} = 3E{-}5\ M\ s^{-1}$, $k_{Glu\text{-}spp} = 7.0866E{-}2\ s^{-1}$
2. $J_{Glu\text{-}spp} = 6E{-}5\ M\ s^{-1}$, $k_{Glu\text{-}spp} = 8.2539E{-}2\ s^{-1}$
3. $J_{Glu\text{-}spp} = 8E{-}5\ M\ s^{-1}$, $k_{Glu\text{-}spp} = 9.0321E{-}2\ s^{-1}$
Substance Flow in Sinusoid
1. $k_e = 0.5$
2. $k_e = 0.8$
3. $k_e = 1.0$
4. $k_e = 1.2$
5. $k_e = 1.6$

GOT, glutamate:oxaloacetate; GDH, glutamate dehydrogenase; GAT, arginine:glycine amidinotransferase; GAMT, guanidinoacetate methyltransferase; OTL, ornithine-citrulline translocase; GTL, glutamate translocase; GATL, glutamate-aspartate translocase; $NH4_4^+$-tp, ammonia transporter; Glu-tp, glutamate transporter; Gln-tp, glutamine transporter in mitochondrial membrane; Urea-tp, urea transporter. The entity abbreviation may be used with an index that represents the location of the entity. The indices *c*, *m*, and *s* indicate the cytoplasm, mitochondria, and sinusoid, respectively.

REVIEW

Informatics for peptide retention properties in proteomic LC-MS

Kosaku Shinoda[1, 2], Masahiro Sugimoto[1, 3], Masaru Tomita[1, 2] and Yasushi Ishihama[1, 4]

[1] Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata, Japan
[2] Human Metabolome Technologies, Tsuruoka, Yamagata, Japan
[3] Bioinformatics Department, Mitsubishi Space Software, Amagasaki, Hyogo, Japan
[4] PRESTO, Japan Science and Technology Agency, Tokyo, Japan

Retention times in HPLC yield valuable information for the identification of various analytes and the prediction of peptide retention is useful for the identification of peptides/proteins in LC-MS-based proteomics. Informatics methods such as artificial neural networks and support vector machines capable of solving nonlinear problems made possible the accurate modeling of quantitative structure-retention relationships of peptides (including large polymers) up to 5 kDa to which classical linear models cannot be applied, as well as the proteome-wide prediction of peptide retention. Proteome-wide retention prediction and accurate mass-information facilitate the identification of peptides in complex proteomic samples. In this review, we address recent developments in solid informatics methods and their application to peptide-retention properties in 'bottom-up' shotgun proteomics. We also describe future prospects for the standardization and application of retention times.

Received: July 13, 2007
Revised: October 30, 2007
Accepted: November 1, 2007

Keywords:
Bioinformatics / Liquid chromatography-tandem mass spectrometry / Neural networks / Peptide / QSRR

1 Introduction

Liquid chromatography-mass spectrometry (LC-MS) is a powerful tool for the separation and identification of peptides in proteomics studies. 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 render accurate peptide/protein identification challenging. Consequently, proteome coverage remains limited. As the chromatographic retention times of peptides depend on their amino acid sequences, their retention times (http://iupac.org/goldbook/R05364.pdf) complement the information provided by MS and enhance their identification. Efforts to predict the chromatographic behavior of peptides span the last 50 years. In 1951, Knight [1] and Pardee [2] showed that in paper chromatography, synthetic peptide retardation factors could be predicted with some accuracy. More recently, the prediction of peptide retention times in RP [3–5] and normal-phase LC [6, 7] was reported. Most of these works 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 linearly dependent on their amino acid composition holds up fairly well for small peptides (up to 15–20 residues), but is

Correspondence: Dr. Yasushi Ishihama, Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017, Japan
E-mail: y-ishi@ttck.keio.ac.jp
Fax: +81-235-29-0536

Abbreviations: ANN, artificial neural networks; **GA**, genetic algorithms; **NET**, normalized elution time; **SMLR**, stepwise multiple linear regressions; **SVM**, support vector machines; **SVR**, support vector regressions; **QSRR**, quantitative structure-retention relationships

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

inadequate for proteomic applications, e.g. those that involve tryptic peptides, where the practical upper limit can exceed 50 amino acid residues [8, 24]. Furthermore, with the retention coefficient approach, isomeric peptides are predicted to elute at the same time, which, in fact, is not the case [9–11]. Another prediction method is based on machine learning methods such as artificial neural networks (ANN) and support vector machines (SVM). Machine-learning techniques capable of solving nonlinear problems [12-23] have been used to model the quantitative structure-retention relationships (QSRR) of various analytes in liquid chromatography [12, 21, 23]. In 2003, Petritis *et al.* [8] introduced an ANN-based method for predicting peptide retention times that was originally based on amino acid composition. Later they extended it to include partial amino acid sequence information [24]. Shinoda *et al.* [25] combined ANN and stepwise multiple linear regressions (SMLR) to predict peptide-retention times based on selected amino acid descriptors with statistically significant effects on LC retentions. Liu *et al.* [26] applied an SVM to develop predictive models between the retention factor (log *k*, http://iupac.org/goldbook/R05359.pdf) and seven peptide molecular constitutional and topological descriptors. These reports confirmed the usefulness of machine learning in peptide-retention predictions especially for longer peptides and several papers applied these techniques to peptide/protein identifications. However, machine learning involves both use and abuse in each step of model development, performance assessment, and application.

This article reviews the strategies, current progress, and underlying difficulties involved in the application of machine-learning methods to the prediction of peptide retentions and examines their application to peptide identification in proteomic studies.

2 Descriptors for peptides

Improving the peptide-retention time prediction in HPLC requires an understanding of the various factors affecting peptide retention behaviors. These factors have been thoroughly investigated [24], and it is now widely accepted that the retention behavior of peptides in HPLC is governed by (i) the amino acid composition [3–5], (ii) the peptide length (or mass) [3, 27, 28], and (iii) sequence-dependent effects [29–40] that can be further divided into nearest-neighbor and conformation effects, where the former are defined as amino acid sequence-dependent but independent of peptide conformation [40]. Krokhin *et al.* [41] applied 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. Using SVM, Liu *et al.* [26] adopted seven peptide molecular constitutional and topological descriptors (*i.e.* number of single bonds, number of rings, etc.) to predict the retention factors (log *k*). Kaliszan and coworkers [42, 43] used QSRR to predict peptide-retention times. Descriptors to derive the necessary QSRR included the logarithm of (i) the sum of the retention times of the amino acids that make up the peptides, (ii) the van der Waals volume of the peptide, and (iii) the peptide-calculated 1-octanol-water partition coefficient. Makrodimitris *et al.* [44] applied a mesoscopic simulation using Langevin dipoles on a lattice with calculated solute partial charges to estimate the free energies of the adsorption of peptides in RP chromatography. Their method is efficient and yields quantitative predictions of retention orders of peptides covering a wide range of structures. Petritis *et al.* [24] investigated several peptide descriptors such as peptide length, sequence, hydrophobicity/hydrophobic moment, and nearest-neighbor amino acid, as well as peptide-predicted structural configurations (*i.e.* helix, sheet, coil). They developed several ANN models with various combinations of these descriptors and empirically assessed the significance of tested descriptors. They found that ANN with a 1052-24-1 architecture, whose input layer consists of encoded peptide sequence information (21 * 25 * 2, amino acids, maximum length, and C/N termini, respectively), peptide length, and hydrophobic moments, yielded the best prediction accuracy.

As outlined above, a number of descriptors have been introduced to represent a peptide; most reported studies typically use only a portion of these descriptors. In other omics applications such as DNA microarray, the selection of a proper subset of descriptors is useful for improving the performance of machine learning methods [45–50]. Moreover, the indiscriminate use of existing descriptors, particularly of overlapping and redundant descriptors, may introduce over-fitting for a particular subset of observable data and deteriorate the versatility of the method. Therefore, there is a need to explore varied combinations of descriptors and to select more optimal sets of descriptors for more cases. This process should not require manual efforts by experts with a deep understanding but rather, it should make use of automatic feature selection methods. For example, Shinoda *et al.* [25] utilized SMLR to select 16 significant descriptors from 20 amino acids to develop ANN while Tham *et al.* [23] applied genetic algorithms (GA) to select molecular descriptors of retention times in RP-HPLC. Efforts have also been directed at improving the efficiency and speed of feature selection methods [51] that will facilitate their more extensive application. Thus, it may be necessary to introduce new descriptors for models that have been described by overlapping and redundant descriptors.

3 Machine-learning methods

Below we describe the concepts and characteristics of representative machine-learning methods including ANN, SVM, and GA. Freely and commercially available solutions of these methods have been reported by Berrueta *et al.* [52] and are listed in Table 1. The characteristics of each method are summarized in Table 2.

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 1. Websites that contain downloadable codes of machine learning methods

	URL	License	Platform
ANN			
Libneural	http://ieee.uow.edu.au/~daniel/software/libneural/	Free (LGPL)	UNIX (GNU/Linux, FreeBSD, NetBSD or OpenBSD) and Cygwin
FANN	http://leenissen.dk/fann/	Free (LGPL)	UNIX/Windows
Weka	http://www.cs.waikato.ac.nz/ml/weka/	Free (GPL)	Cross-platform (Java)
NeuralWorks Predict	http://www.neuralware.com/products.jsp	Commercial	Windows/UNIX
NeuroShell Predictor	http://www.mbaware.com/neurpred.html	Commercial	Windows 95-XP
BrainMaker	http://www.calsci.com/	Commercial	Windows XP, 2000 and Me
JMP	http://www.jmp.com/software/jmp.shtml	Commercial	Windows/MacOS X/Linux
SVM			
SVM light	http://svmlight.joachims.org/	Free for non-commercial use	PowerPC Mac/UNIX/Windows
LIBSVM	http://www.csie.ntu.edu.tw/~cjlin/libsvm/	Free (the modified BSD license)	Windows/UNIX
mySVM	http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/	Free for non-commercial use	Windows/UNIX
BSVM	http://www.csie.ntu.edu.tw/~cjlin/bsvm/	Free for non-commercial use	Windows/UNIX
Weka	http://www.cs.waikato.ac.nz/ml/weka/	Free (GPL)	Cross-platform (Java)
MATLAB SVM Toolbox	http://theoval.sys.uea.ac.uk/svm/toolbox/	GPL (Matlab is Commercial)	Windows/UNIX/Mac OS X
GA			
AI::Genetic (CPAN module)	http://search.cpan.org	Free (GPL)	UNIX (Solaris, Linux, FreeBSD, NetBSD or OpenBSD) and Cygwin
GAlib	http://lancet.mit.edu/ga/	Free (GPL)	UNIX (Linux, MacOSX, SGI, Sun etc)/Windows/Mac OS
genalgy	http://hobbiton.thisside.net/genetic/	Free	Python code
JGAP	http://jgap.sourceforge.net/	Free (LPL or MPL)	Cross-platform (Java)
Weka	http://www.cs.waikato.ac.nz/ml/weka/	Free (GPL)	Cross-platform (Java)
Genetic Algorithm and Direct Search Toolbox Matlab	http://www.mathworks.com/products/gads/	Commercial	Windows/UNIX/Mac OS X

Table 2. Brief comparisons of machine learning methods described in this review

	Accuracy for nonlinear problems	Model interpretability	Preferable dataset size	Generalization ability	Possibility of over-fitting
MLR	–	++	$> X^{a)}+1$	Low	None
ANN	++	+	$> X*5.0^{b)}$	High	High
SVM (SVR)	++	–	$> X*2.0^{c)}$	High	Low

a) X is the number of variable.
b) Rough requirement. The preferable size strongly depends on the number of hidden nodes.
c) Rough requirement. The preferable size depends on the kernel function.

3.1 Artificial neural networks

An ANN is a generic designation for connectionist-approach-based data modeling tools inspired by the biological nervous system. ANN can be used to detect underlying relationships between inputs and outputs or to find patterns in data. Compared to classical statistical methods, ANN-based approaches offer advantages that include a capacity to self-learn and to model complex data without the need for a detailed understanding of the underlying phenomena.

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Among various types of ANN, a multi-layer perceptron is the most common algorithm; it has been widely used for peptide retention predictions [8, 24, 25, 53, 54]. It is composed of a large number of neurons, nodes, or processing elements organized into a sequence of layers. As shown in Fig. 1, nodes in any layer can be fully or partially connected to nodes of a succeeding layer; each hidden or output node receives signals in parallel. The input signal to a node is modulated by a weight (w) along each link between nodes. The net input to a node is thus a function of all signals to the node and all of its associated weights. For example, the net input for a node j is given by:

$$net_j = \sum_i w_{ji} O_i \quad (1)$$

where i represents nodes in the previous layer, w_{ji} is the weight associated with the connection from node i to node j, and O_i is the output of node i. The process of adapting the weights to an optimal set of values is called "training" the neural network. For this, several training algorithms are available; the back-propagation (backwards propagation of errors) algorithm illustrated in Fig. 2 is the most popular [55]. The net inputs were transferred to the neuron using a transfer function. Several transfer functions are available, satisfying a requirement of differentiability set by the back-propagation algorithm. The most popular is the logistic function given by:

$$O_i = \frac{1}{1 + e^{-net_i}} \quad (2)$$

Overall, the structure of a multi-layer perceptron contains at least three layers, *i.e.* an input layer with one node for each variable in a data vector and an output layer consisting of one node for each variable to be investigated. Additionally, one or more hidden layers can be added between the input and output. Funahashi [56] previously demonstrated that a single

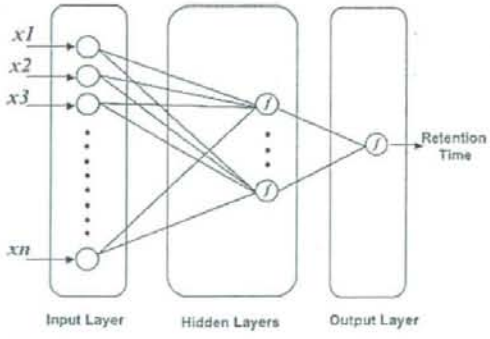

Figure 1. Schematic representation of the artificial neural network architecture. The circles represent input vectors. The small black circles show continuance.

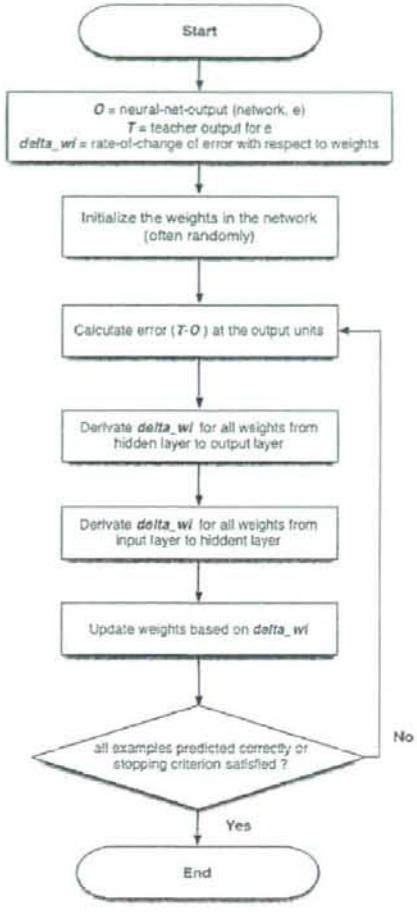

Figure 2. Algorithmic representation of back propagation (back propagation of errors).

hidden layer could approximate any function. Without hidden layers, a neural network with logistic transfer function is identical to logistic regression widely used in statistical modeling. In essence, the application of these equations to nodes in the hidden and output layers allows these ANN to perform multivariate nonlinear regression using a logistic function. Due to the parallel processing of nodes within each layer, these ANN can learn multivariate nonlinear functions.

3.2 Support vector machine

The support vector machine developed by Vapnik *et al.* [57–60] as a novel type of machine-learning method is gaining popularity due to its many attractive features and promising

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

empirical performance. Compared to traditional ANN, SVM features the following prominent advantages: (i) a strong theoretical background provides SVM with a high generalization capability and can avoid local minima, *i.e.* it has the ability to accurately predict for new data, (ii) SVM always reaches a solution that can be quickly obtained by a standard linear optimization algorithm (quadratic programming), (iii) SVM does not need to determine network topology in advance, rather, it can be automatically obtained at the end of the training process, and (iv) SVM builds a result based on a sparse subset of training samples, thereby reducing the workload. Originally, SVM was developed to solve pattern recognition problems; it is now used for microarray gene expression classification [61], protein folding recognition [62], protein structural class prediction [63], identification of protein cleavage sites, and other pharmaceutical data analyses [61, 64]. Vapnik [58] extended this algorithm for solving regression problems by choosing a suitable cost function (ε-insensitive loss function) that facilitates the acquisition of a sparse set of support vectors (support vector regressions, or SVR). Although SVR has been used in the prediction of chromatographic behavior such as log *k* of peptides [26] and of protein retentions in anion-exchange- [65] and hydrophobic interaction chromatography [66], proteome-wide applications of SVR have just begun. The basic concept of SVM has been described and illustrated in clearly understandable terms by Noble [67].

3.3 Genetic algorithm

The GA [68] is an algorithm based on evolutionary computation and survival of the fittest; it is often applied to optimization problems such as optimizing the free variables in a hypothesis function [69]. Solutions to problems are coded as genes (= abstract representations of variables to be optimized) in each individual (= candidate solutions). Traditionally, genes are represented in binary form as strings of 0 and 1; other encodings (*e.g.* real number) are also possible. At first, a number of individuals are initialized with randomly generated genes to form a "population". The fitness of each individual in the population is evaluated by a user-specified fitness function; multiple individuals are stochastically selected from the current population based on their fitness and modified (recombined by crossover operation and randomly mutated) to form a new population. This iterative process, called a "generation", incrementally refines the best solution, the fittest individuals. The GA performs a global search that avoids local minima; many parameters can be estimated simultaneously [70]. In terms of practical applications, the GA was used to select molecular descriptors of retention times in RP-HPLC [23], to optimize numeric parameters of normalization functions [8], and to optimize scoring functions for protein identification [23]. The basic concept and theory as well as influences of parameters empirically tuned by users of GA are described by Leardi [71].

4 Assessment of the performance of predictive models

In using machine learning, a fundamental question is how best to assess the performance of predictors [77]. One way is to obtain a test set of further observations from the same population and to compare the observed values in this set with their predictions from the model using a single criterion measure such as the sum of squared errors. When no test set is available, we need to base assessments on training-set data only. The simplest way is resubstitution, *i.e.* comparing predictions for individual data in the training set with their counterparts. However, this will give optimistic assessments, because predictors perforce find predictive values closest to the values in the training set. Such close matching will not occur for independently gathered data. A favored alternative is cross-validation [72, 73]. Here the training data are divided into *g* equal-sized groups and *g* separate operations are conducted. Each group is omitted in turn from the data, the model is fitted to the remaining $(g - 1)$ groups, and the predictions are obtained for the omitted group. This yields n (= the number of individual data in the whole training set) predictions, none of which used the corresponding training data as part of the modeling stage. Therefore, the performance assessments formed from these predictions should not be optimistically biased. As the number of individuals in each omitted group is $k = n/g$, this method of assessment is termed *leave-k-out*. Leave-*k*-out cross-validation was popularized in the bioinformatics and chemoinformatics areas, where it is often termed "*g*-fold" cross-validation. Theoretical and computational investigations have been conducted into the influence of *k* on the results. Shao [74] established that consistency improves as *k* increases and Altman and Leger [75] reached a similar conclusion with respect to asymptotic optimality. A complication arises because there are $n!/(g!(k!)^g)$ ways of dividing the training set into *g* groups each with size *k*, and different partitions may yield different performance assessments. One solution is to average criterion measures over different partitions to arrive at an overall assessment.

Machine learning-based predictive models often depend on parameters that can only be optimized (estimated) through data-based inspection. For example, SVM reformulates the model in terms of a user-specified parameter ε that controls how closely the function will fit the training data but requires a prior determination of ε before fitting the model [76]. ANN has more varied empirically tuned parameters such as the learning rate, momentum, and number of hidden nodes; these parameters must be selected in anticipation of learning. The GA is more complicated; various parameters such as the number of generations and populations, and the mutation- and crossover rate should be determined empirically. Such parameter selection can be performed using cross-validation. For example, the number of hidden nodes to include in the ANN model can be chosen as the number that yields the lowest predictive error when successively fitting 1, 2, 3, 4 nodes, and so on. This process is called tuning

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[77]. Although assessment of the performance of tuned models on test data is the best approach, what is to be done in the absence of a test set? The sum of squared errors for the chosen model is clearly an optimistically biased assessment because the model has been chosen to give the lowest errors on the training data. For unbiased assessment, we need a second layer of cross-validation: leave out each group of individuals in turn, use cross-validation on the remaining individuals to both tune and fit the model, and then make predictions for the omitted individuals using the fitted model. This process is called *two-deep* as opposed to the *one-deep* cross-validation described earlier. The necessity for two-deep cross-validation has been stressed by Ganeshanandam and Krzanowski [78] whenever predictive models are constructed by optimizing cross-validation error rates, and by Krzanowski [79] whenever the selection of variables is based on cross-validated error rates. Despite such warnings, there is often still a reluctance to use two-deep validations. Many papers on the application of machine learning in proteomic studies continued to perform one-deep assessment of error of a predictive model (Table 3); a few appropriately used two-deep cross-validations [21, 25]. Representative results from two-deep assessments of ANN are shown in Fig. 3.

The aim of cross-validation is to mimic the prediction for *future* individuals from the population. This will be achieved if the training data fully represent the sample space and each omitted individual can lie anywhere in this space. Large samples and small dimensionality generally satisfy these requirements. With small samples and high dimensionality, the training data are likely to fall in a very small fraction of the sample space (the "curse of dimensionality"; [80]), and any omitted group from the training set will only come from this restricted area. Cross-validation may therefore fall far short of replicating the conditions of a test set, consequently, as dimensionality increases, the method may become less reliable. Therefore, as stated above, there is a need to explore different combinations of descriptors and to discriminate more optimal subsets of descriptors; this can be done by using feature selection methods [45, 46, 49].

5 Application to peptide and protein identifications

There are several studies on the prediction of the LC retention time of tryptic peptides for protein identification. In 2002, Palmblad *et al.* [81] first showed that retention time prediction could be combined with PMF to improve protein identification in proteomic experiments; however, their peptide retention time prediction error was high, presumably

Table 3. Performance of machine learning methods for predicting peptide retentions as reported in the literature

Machine learning	Peptide descriptors	Peptide types	Number of peptides in dataset	Validation method	Reported prediction accuracy		Ref.
					Correlation coefficient (R) or R-squared	Error rate	
ANN/ SMLR	Amino acid composition	LysC-digested peptides	834	Ten-fold two-deep CV	0.928 (R-squared)	<4%	[25]
ANN	Amino acid composition	Tryptic peptides	7080	One-deep CV	-	<3%	[8]
ANN	Peptide sequence, length, hydrophobic moment	Tryptic peptides	345914	One-deep CV	0.967 (R)	<3%	[24]
SVM	Average complementary information content, relative number of single bonds, relative number of S and N, average information content (order 0/2) and number of rings	Enzymatic digestion (trypsin and lysyl endopeptidase) of purified proteins	75	Independent evaluation	0.9801 (R)	0.1523 (in log of retention factor)	[26]
SVM	Number of histidines, histidine pairs and arginines/isoelectric point of peptides sequence	Synthesized peptides	Several hundreds	Bootstrap	0.85 (R-squared)	-	[108]

All data and results shown were collected from the original papers. The reported prediction performances must be interpreted cautiously because they are dependent on factors such as the datasets used and the choice of parameters.

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

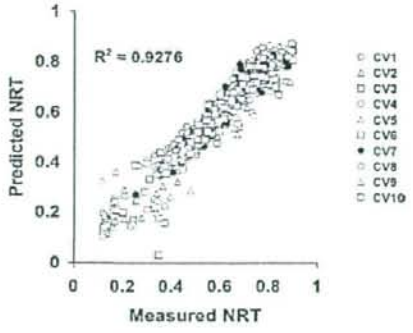

Figure 3. Scattergram of the correlation between experimentally measured and predicted normalized LC retention times (NRT) for all 834 peptides through ten-fold two-deep cross-validations (CV) from [25]; permission was obtained from the authors. NRT was predicted using ANN. Parameters for ANN (training ratio, momentum, and random numbers for initial ANN weights) were tuned through ten-fold CVs using 90% of 834 peptides and performance of the tuned model was evaluated using the remaining 10%.

because of limitations of the retention coefficient approach they used. Smith and co-workers [82–85] reported an accurate mass and time (AMT) tag proteomics approach that applies accurate mass measurements in conjunction with observed peptide-retention time information to identify peptides more confidently. Le Bihan *et al.* [86] used peptide-retention time prediction parameters to build a model for predicting peptides that are likely to be observable by LC-MS/MS; their model was employed for the targeted MS identification of low-abundance proteins in complex protein samples. Kawakami *et al.* [87] developed a program that validates peptide assignments based on the correlation between the measured and predicted LC retention time of each peptide. Norbeck *et al.* [54] 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. When peptide-retention time prediction was combined with peptide/protein identification programs such as SEQUEST and MASCOT in various applications, the number of false-positive identifications could be decreased [88–90].

In using peptide retention information for identification, comparing multiple LC-MS/MS runs or matching observed and predicted retention times is challenging because small changes in the split ratio, column lengths, column packings, void volumes, etc. unavoidably lead to some retention-time variability. In addition, noncontiguous retention data obtained with different LC-MS systems or by different laboratories must be aligned to confirm and utilize established proteome data. A widely used approach to the chromatographic-alignment problem fits a piece-wise linear function to maximize the correlation between the samples. Methods of this kind are often characterized as correlation optimized warping (COW) [91] and several derivative methods were investigated [92]. In principle, this approach can be extended to aligning multi-dimensional data. However, the handling of proteomic data is extremely difficult because the data are typically characterized by a very large input dimension (*i.e.* tryptic peptides). Thus, more sophisticated alignment algorithms are needed to extract higher quality information from large-scale LC-MS-based experiments. A number of approaches has been developed and used in high-throughput proteomic applications that rely on combined results obtained from different experimental platforms. For example, in the AMT approach [53, 54, 83, 93, 94], results from different LC-MS or MS/MS data sets are combined by finding the transformation functions of mass and retention times that are required to remove variability in mass and retention-time measurements between analyses. An alternative approach by Radulovic *et al.* [95] developed a software suite that bins peaks from MS scans by m/z bins and uses signal-processing algorithms to discover peaks in the chromatographic data, which contain pixels for identified peaks. Pamphlets from different experiments are aligned by using a 2-D smoothing spline function in the m/z and time dimensions to correct for m/z and time drift. Listgarten *et al.* [96] described a method to concurrently align multiple datasets by using a continuous profile model (CPM), a generative model in which each observed time series is a non-uniformly sub-sampled version of a single latent trace, to which local scale transformations are applied. Another proposed pipeline utilizes dynamic programming (DP) [97, 98]. Prakash *et al.* [97] performed DP-based alignment using a score that assumes the similarity of intensity profiles of mass spectra in different LC-MS analyses. Multiple analyses are combined in a progressive strategy of aligning and merging datasets based on similarity.

Machine learning is also applied to develop an "intelligent" system for comparing a large number of LC/MS experiments. Petritis *et al.* [8] introduced GA to optimize the normalization function for peptide retention times. The GA was applied to >50 000 (9121 distinct) peptides identified from 687 LC/MS/MS analyses to establish a common timeline so that the same peptides' variances of NET (normalized between 0 and 1) across the different separations were minimized. The GA optimized two variables of the linear normalization function for each separation to reduce the variance function of specific peptides, *i.e.* the regressed retention times for each separation. While this generated excellent results, this normalization approach became time-prohibitive as the number of peptides used increased significantly due to the many generations (iterations) required to align all analyses [24]. To remove this limitation, Strittmatter *et al.* [82] regressed observed retention times of confidently identified peptides to predicted NET of the sequences using a quadratic function for each LC-MS run. The obtained quadratic equations were used to convert observed retention times to

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

observed NET and all LC-MS runs could be compared on scales of the NET. To apply their methodology across different laboratories, at least the following three requirements must be satisfied (i) NET predictors should be available, (ii) analytical columns as well as the mobile phase should be identical, and (iii) not only the variables but also the functions for conversion should be optimized for each laboratory. However, even if these requirements were met, the above approach cannot be applied to cases where retention time reversal [99] occurs between different gradients due to the sensitivity of peptide retentions to changes in the concentration of organic solvents.

As an alternative approach, we investigated whether *log* k_0 (logarithm of retention factor for a given organic solvent) of the linear solvent strength (LSS) theory [99] can be utilized as a peptide-specific "universal" retention index that is independent from LC gradients and depends solely on the constituent of the mobile phase and columns. We introduced a machine learning scheme to optimize the transformation function between retention times and *log* k_0. With the optimized function, peptide-retention data obtained from different gradients can be compared on scales of *log* k_0 and used among different laboratories performing multiple experiments including retention-time reversal [100, 101] (Shinoda *et al.*, submitted).

Each of these approaches has its own computational requirements and implicit challenges based on how the data are preprocessed. Our laboratory continues to study the extensive application of new machine learning techniques not only to predictions but also to the "standardization" of peptide-retention times.

6 Conclusions and perspectives

Machine-learning methods have been explored as valuable tools for predicting peptide retention times. A number of studies have consistently demonstrated the usefulness of these methods for predicting peptide retentions and their applicability to peptide identifications in proteomic studies. Because of their *m/z*-independent nature, these methods are useful for studying complex proteomic samples that cannot be completely analyzed by current protocols. Furthermore, they can be applied to the expression profiling of a substantial number of unknown ORF in many of the currently completed genomes [102]. Existing algorithms can be improved and new algorithms may be introduced to enhance the performance and accuracy of machine-learning methods.

The extensive application of new machine-learning techniques not only to predictions but also to the "normalization" and "standardization" of peptide-retention times is desirable to accelerate the inter-laboratory use of retention time predictors and the utilization of published proteomic LC-MS data. When assessing the performance of the new predictive models, two-deep cross-validations are preferable to avoid optimistically biased results and over-fitting.

In this review we focused on the application of peptide retentions for identifications; however, retention parameters can also be utilized for quantifications. A proteotypic peptide probe (P3), that is, a frequently observable peptide that uniquely identifies a specific protein or protein isoform, has come into use as a target for high-throughput protein quantification. The accumulation of large-scale P3 datasets [103, 104] has accelerated the transition in current proteomics from a discovery to a scoring phase [105]. The normalization of obtained P3 retention-time data using the above machine-learning techniques and the application of targeted MS scans to the retention-time range are possible applications.

Models for predicting the retention times of peptides with PTM (*e.g.* phosphopeptides) are needed. Recent studies have shown that contrary to general expectations, synthetic phosphopeptides were eluted at almost the same position, or slightly more slowly, than the corresponding nonphosphopeptides from a C18 column [106, 107]. The further acquisition of this type of data and the machine learning techniques described herein will facilitate the accurate prediction of phosphopeptide retention times and false-positive identifications may be excluded in current phospho-proteomic experiments on the basis of distinctive retention times.

Machine-learning approaches will be applied to predict the retention time of peptides separated by other chromatography modes such as ion-exchange chromatography and hydrophilic interaction chromatography for further quality assurance. This will add another dimension of confidence and will be especially useful to research groups that use on-line (*e.g.* MudPIT) or off-line 2-D chromatography for peptide separation/fractionation.

Due to the rapid progress in proteomics, the prediction capability of machine-learning methods is further enhanced by the increasing availability of large-scale data that generates variability of peptide sequences and more extensive knowledge about primary sequences, PTM, and structures that define the chromatographic behavior of peptides.

This work was supported by research funds from the Yamagata prefectural government and Tsuruoka city.

The authors have declared no conflict of interest.

7 References

[1] Knight, C. A., Paper chromatography of some lower peptides. *J. Biol. Chem.* 1951, *190*, 753–756.

[2] Pardee, A. B., Calculations on paper chromatography of peptides. *J. Biol. Chem.* 1951, *190*, 757–762.

[3] Mant, C. T., Burke, T. W., Black, J. A., Hodges, R. S., Effect of peptide chain length on peptide retention behaviour in reversed-phase chromatography. *J. Chromatogr.* 1988, *458*, 193–205.

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[4] Meek, J. L., Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. *Proc. Natl. Acad. Sci. USA* 1980, *77*, 1632–1636.

[5] Wilce, M. C., Aguilar, M. I., Hearn, M. T., High-performance liquid chromatography of amino acids, peptides and proteins. CXXII. Application of experimentally derived retention coefficients to the prediction of peptide retention times: studies with myohemerythrin. *J. Chromatogr.* 1993, *632*, 11–18.

[6] Yoshida, T., Calculation of peptide retention coefficients in normal-phase liquid chromatography. *J. Chromatogr. A* 1998, *808*, 105–112.

[7] Yoshida, T., Okada, T., Prediction of peptide retention times in normal-phase liquid chromatography with only a single gradient run. *J. Chromatogr. A* 1999, *841*, 19–32.

[8] Petritis, K., Kangas, L. J., Ferguson, P. L., Anderson, G. A. *et al.*, Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. *Anal. Chem.* 2003, *75*, 1039–1048.

[9] Hearn, M. T., Aguilar, M. I., High-performance liquid chromatography of amino acids, peptides and proteins. LXIX. Evaluation of retention and bandwidth relationships of myosin-related peptides separated by gradient elution reversed-phase high-performance liquid chromatography. *J. Chromatogr.* 1987, *392*, 33–49.

[10] Petritis, K., Brussaux, S., Guenu, S., Elfakire, C., Dreux, M., Ion-pair reversed-phase liquid chromatography-electrospray mass spectrometry for the analysis of underivatized small peptides. *J. Chromatogr. A* 2002, *957*, 173–185.

[11] Terabe, S., Konaka, R., Inouye, K., Separation of some polypeptide hormones by high-performance liquid chromatography. *J. Chromatogr.* 1979, *172*, 163–177.

[12] Agatonovic-Kustrin, S., Zecevic, M., Zivanovic, L., Use of ANN modelling in structure–retention relationships of diuretics in RP-HPLC. *J. Pharm. Biomed. Anal.* 1999, *21*, 95–103.

[13] Aires-de-Sousa, J., Hemmer, M. C., Gasteiger, J., Prediction of 1H NMR chemical shifts using neural networks. *Anal. Chem.* 2002, *74*, 80–90.

[14] Havel, J., Breadmore, M., Macka, M., Haddad, P. R., Artificial neural networks for computer-aided modelling and optimisation in micellar electrokinetic chromatography. *J. Chromatogr. A* 1999, *850*, 345–353.

[15] Jalali-Heravi, M., Fatemi, M. H., Prediction of thermal conductivity detection response factors using an artificial neural network. *J. Chromatogr. A* 2000, *897*, 227–235.

[16] Jalali-Heravi, M., Fatemi, M. H., Artificial neural network modeling of Kovats retention indices for noncyclic and monocyclic terpenes. *J. Chromatogr. A* 2001, *915*, 177–183.

[17] Jalali-Heravi, M., Garkani-Nejad, Z., Prediction of electrophoretic mobilities of sulfonamides in capillary zone electrophoresis using artificial neural networks. *J. Chromatogr. A* 2001, *927*, 211–218.

[18] Jalali-Heravi, M., Garkani-Nejad, Z., Prediction of electrophoretic mobilities of alkyl- and alkenylpyridines in capillary electrophoresis using artificial neural networks. *J. Chromatogr. A* 2002, *971*, 207–215.

[19] Malovana, S., Frias-Garcia, S.,Havel, J., Artificial neural networks for modeling electrophoretic mobilities of inorganic cations and organic cationic oximes used as antidote contra nerve paralytic chemical weapons. *Electrophoresis* 2002, *23*, 1815–1821.

[20] Muzikar, M., Havel, J., Macka, M., Capillary electrophoresis determinations of trace concentrations of inorganic ions in large excess of chloride: soft modelling using artificial neural networks for optimisation of electrolyte composition. *Electrophoresis* 2003, *24*, 2252–2258.

[21] Ruggieri, F., D'Archivio, A. A., Carlucci, G., Mazzeo, P., Application of artificial neural networks for prediction of retention factors of triazine herbicides in reversed-phase liquid chromatography. *J. Chromatogr. A* 2005, *1076*, 163–169.

[22] Sugimoto, M., Kikuchi, S., Arita, M., Soga, T. *et al.*, Large-scale prediction of cationic metabolite identity and migration time in capillary electrophoresis mass spectrometry using artificial neural networks. *Anal. Chem.* 2005, *77*, 78–84.

[23] Tham, S. Y., Agatonovic-Kustrin, S., Application of the artificial neural network in quantitative structure-gradient elution retention relationship of phenylthiocarbamyl amino acids derivatives. *J. Pharm. Biomed. Anal.* 2002, *28*, 581–590.

[24] Petritis, K., Kangas, L. J., Yan, B., Monroe, M. E. *et al.*, Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Anal. Chem.* 2006, *78*, 5026–5039.

[25] Shinoda, K., Sugimoto, M., Yachie, N., Sugiyama, N. *et al.*, Prediction of liquid chromatographic retention times of peptides generated by protease digestion of the *Escherichia coli* proteome using artificial neural networks. *J. Proteome Res.* 2006, *5*, 3312–3317.

[26] Liu, H. X., Xue, C. X., Zhang, R. S., Yao, X. J. *et al.*, Quantitative prediction of logk of peptides in high-performance liquid chromatography based on molecular descriptors by using the heuristic method and support vector machine. *J. Chem. Inf. Comput. Sci.* 2004, *44*, 1979–1986.

[27] O'Hare, M. J., Nice, E. C., Hydrophobic high-performance liquid chromatography of hormonal polypeptides and proteins on alkylsilane-bonded silica. *J. Chromatogr.* 1979, *171*, 209–226.

[28] Su, S.-J., Grego, B., Niven, B., Hearn, M. T. W., Analysis of group retention contributions for peptides separated by reversed phase high performance liquid chromatography. *J. Liquid Chromatogr. Rel. Technol.* 1981, *4*, 1745–1764.

[29] Blondelle, S. E., Buttner, K., Houghten, R. A., Evaluation of peptide-peptide interactions using reversed-phase high-performance liquid chromatography. *J. Chromatogr.* 1992, *625*, 199–206.

[30] Blondelle, S. E., Ostresh, J. M., Houghten, R. A., Perez-Paya, E., Induced conformational states of amphipathic peptides in aqueous/lipid environments. *Biophys. J.* 1995, *68*, 351–359.

[31] Buttner, K., Pinilla, C., Appel, J. R., Houghten, R. A., Anomalous reversed-phase high-performance liquid chromatographic behavior of synthetic peptides related to antigenic helper T cell sites. *J. Chromatogr.* 1992, *625*, 191–198.

[32] Chen, Y., Mant, C. T., Hodges, R. S., Temperature selectivity effects in reversed-phase liquid chromatography due to conformation differences between helical and non-helical peptides. *J. Chromatogr. A* 2003, *1010*, 45–61.

[33] Rothemund, S., Krause, E., Beyermann, M., Dathe, M. *et al.*, Recognition of alpha-helical peptide structures using high-

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

performance liquid chromatographic retention data for D-amino acid analogues: influence of peptide amphipathicity and of stationary phase hydrophobicity. *J. Chromatogr. A* 1995, *689*, 219–226.

[34] Sereda, T. J., Mant, C. T., Sonnichsen, F. D., Hodges, R. S., Reversed-phase chromatography of synthetic amphipathic alpha-helical peptides as a model for ligand/receptor interactions. Effect of changing hydrophobic environment on the relative hydrophilicity/hydrophobicity of amino acid side-chains. *J. Chromatogr. A* 1994, *676*, 139–153.

[35] Steer, D. L., Thompson, P. E., Blondelle, S. E., Houghten, R. A., Aguilar, M. I., Comparison of the binding of alpha-helical and beta-sheet peptides to a hydrophobic surface. *J. Pept. Res.* 1998, *51*, 401–412.

[36] Su, J. Y., Hodges, R. S., Kay, C. M., Effect of chain length on the formation and stability of synthetic alpha-helical coiled coils. *Biochemistry* 1994, *33*, 15501–15510.

[37] Wieprecht, T., Rothemund, S., Bienert, M., Krause, E., Role of helix formation for the retention of peptides in reversed-phase high-performance liquid chromatography. *J. Chromatogr. A* 2001, *912*, 1–12.

[38] Wimley, W. C., Creamer, T. P., White, S. H., Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. *Biochemistry* 1996, *35*, 5109–5124.

[39] Yu, Y. B., Wagschal, K. C., Mant, C. T., Hodges, R. S., Trapping the monomeric alpha-helical state during unfolding of coiled-coils by reversed-phase liquid chromatography. *J. Chromatogr. A* 2000, *890*, 81–94.

[40] Zhou, N. E., Mant, C. T., Hodges, R. S., Effect of preferred binding domains on peptide retention behavior in reversed-phase chromatography: amphipathic alpha-helices. *Pept. Res.* 1990, *3*, 8–20.

[41] Krokhin, O. V., Craig, R., Spicer, V., Ens, W. *et al.*, An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: its application to protein peptide mapping by off-line HPLC-MALDI MS. *Mol. Cell. Proteomics* 2004, *3*, 908–919.

[42] Kaliszan, R., Baczek, T., Cimochowska, A., Juszczyk, P. *et al.*, Prediction of high-performance liquid chromatography retention of peptides with the use of quantitative structure-retention relationships. *Proteomics* 2005, *5*, 409–415.

[43] Baczek, T., Wiczling, P., Marszall, M., Heyden, Y. V., Kaliszan, R., Prediction of peptide retention at different HPLC conditions from multiple linear regression models. *J. Proteome Res.* 2005, *4*, 555–563.

[44] Makrodimitris, K., Fernandez, E. J., Woolf, T. B., O'Connell, J. P., Mesoscopic simulation of adsorption of peptides in a hydrophobic chromatography system. *Anal. Chem.* 2005, *77*, 1243–1252.

[45] Al-Shahib, A., Breitling, R., Gilbert, D., Feature selection and the class imbalance problem in predicting protein function from sequence. *Appl. Bioinformatics* 2005, *4*, 195–203.

[46] Al-Shahib, A., Breitling, R., Gilbert, D., FrankSum: new feature selection method for protein function prediction. *Int. J. Neural Syst.* 2005, *15*, 259–275.

[47] Li, L., Jiang, W., Li, X., Moser, K. L. *et al.*, A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. *Genomics* 2005, *85*, 16–23.

[48] Sindhwani, V., Rakshit, S., Deodhare, D., Erdogmus, D. *et al.*, Feature selection in MLPs and SVMs based on maximum output information. *Neural Networks, IEEE Transactions on* 2004, *15*, 937–948.

[49] Xue, Y., Li, Z. R., Yap, C. W., Sun, L. Z. *et al.*, Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inf. Comput. Sci.* 2004, *44*, 1630–1638.

[50] Zhang, X., Lu, X., Shi, Q., Xu, X. Q. *et al.*, Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics* 2006, *7*, 197.

[51] Furlanello, C., Serafini, M., Merler, S., Jurman, G., An accelerated procedure for recursive feature ranking on microarray data. *Neural Netw.* 2003, *16*, 641–648.

[52] Berrueta, L. A., Alonso-Salces, R. M., Heberger, K., Supervised pattern recognition in food analysis. *J. Chromatogr. A*, in press.

[53] Jaitly, N., Monroe, M. E., Petyuk, V. A., Clauss, T. R. *et al.*, Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Anal. Chem.* 2006, *78*, 7397–7409.

[54] Norbeck, A. D., Monroe, M. E., Adkins, J. N., Anderson, K. K. *et al.*, The utility of accurate mass and LC elution time information in the analysis of complex proteomes. *J. Am. Soc. Mass Spectrom.* 2005, *16*, 1239–1249.

[55] Rumelhart, D. E., Hinton, G. E., Williams, R. J., Learning representations by back-propagating errors. *Nature* 1986, *323*, 533–536.

[56] Funahashi, K., On the approximate realization of continuous mappings by neural networks. *Neural Netw.* 1989, *2*, 183–192.

[57] Vapnik, V. N., *Statistical learning theory*, Wiley, New York 1998.

[58] Vapnik, V., *The nature of statistical learning theory*, Springer-Verlag, New York 1995.

[59] Cortes, C., Vapnik, V., Support-vector networks. *Machine Learning* 1995, *20*, 273–297.

[60] Bernhard, E. B., Isabelle, M. G., Vladimir, N. V., *Proceedings of the fifth annual workshop on Computational learning theory*, ACM Press, Pittsburgh, PA, USA 1992.

[61] Burbidge, R., Trotter, M., Buxton, B., Holden, S., Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* 2001, *26*, 5–14.

[62] Ding, C. H., Dubchak, I., Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* 2001, *17*, 349–358.

[63] Karchin, R., Karplus, K., Haussler, D., Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* 2002, *18*, 147–159.

[64] Czerminski, R., Yasri, A., Hartsough, D., Use of support vector machine in pattern classification: Application to QSAR studies. *Quant. Struct.-Act. Relat.* 2001, *20*, 227–240.

[65] Song, M., Breneman, C. M., Bi, J., Sukumar, N. *et al.*, Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. *J. Chem. Inf. Comput. Sci.* 2002, *42*, 1347–1357.

[66] Ladiwala, A., Xia, F., Luo, Q., Breneman, C. M., Cramer, S. M., Investigation of protein retention and selectivity in HIC sys-

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

tems using quantitative structure retention relationship models. *Biotechnol. Bioeng.* 2006, *93*, 836–850.

[67] Noble, W. S., What is a support vector machine? *Nat. Biotechnol.* 2006, *24*, 1565–1567.

[68] Holland, J. H., *Adaption in natural and artificial systems: An introductory analysis with applications to Biology, control and artificial intelligence. University of Michigan Press* 1975.

[69] Holland, J. H., *Adaptation in natural and artificial systems*, MIT Press, Cambridge, MA, USA 1992.

[70] Goldberg, D. E., *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley Pub. Co., Reading, MA, USA 1989.

[71] Leardi, R., Genetic algorithms in chemistry. *J. Chromatogr. A* 2007, *1158*, 226–233

[72] Stone, M., Cross-validatory choice and assessment of statistical predictions. *Jo. Roy. Statist. Soc. Series B (Methodological)* 1974, *36*, 111–147.

[73] Lachenbruch, P. A., Mickey, M. R., Estimation of error rates in discriminant analysis. *Technometrics* 1968, *10*, 1–11.

[74] Shao, J., Linear model selection by cross-validation. *J. Am. Statist. Assoc.* 1993, *88*, 486–494.

[75] Altman, N., Leger, C., On the optimality of prediction-based selection criteria and the convergence rates of estimators. *J. Roy. Statist. Soc. Series B (Methodological)* 1997, *59*, 205–216.

[76] Witten, I. H., Frank, E., *Data mining: Practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, San Mateo, CA 1999.

[77] Jonathan, P., Krzanowski, W. J., McCarthy, W. V., On the use of cross-validation to assess performance in multivariate prediction. *Statistics Computing* 2000, *10*, 209–229.

[78] Ganeshanandam, S., Krzanowski, W. J., On selecting variables and assessing their performance in linear discriminant analysis. *Aust. J. Stat.* 1989, *31*, 433–447.

[79] Krzanowski, W. J., Selection of variables, and assessment of their performance, in mixed-variable discriminant analysis. *Comput. Statist. Data Analysis* 1995, *19*, 419–431.

[80] Bellman, R., *Adaptive control processes: a guided tour*, Princeton University Press, Princeton 1961.

[81] Palmblad, M., Ramstrom, M., Markides, K. E., Hakansson, P., Bergquist, J., Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry. *Anal. Chem.* 2002, *74*, 5826–5830.

[82] Strittmatter, E. F., Ferguson, P. L., Tang, K., Smith, R. D., Proteome analyses using accurate mass and elution time peptide tags with capillary LC time-of-flight mass spectrometry. *J. Am. Soc. Mass Spectrom.* 2003, *14*, 980–991.

[83] Smith, R. D., Anderson, G. A., Lipton, M. S., Pasa-Tolic, L. *et al.*, An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* 2002, *2*, 513–523.

[84] Lipton, M. S., Pasa-Tolic, L., Anderson, G. A., Anderson, D. J. *et al.*, Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc. Natl. Acad. Sci. USA* 2002, *99*, 11049–11054.

[85] Conrads, T. P., Anderson, G. A., Veenstra, T. D., Pasa-Tolic, L., Smith, R. D., Utility of accurate mass tags for proteome-wide protein identification. *Anal. Chem.* 2000, *72*, 3349–3354.

[86] Le Bihan, T., Robinson, M. D., Stewart, II, Figeys, D., Definition and characterization of a "trypsinosome" from specific peptide characteristics by nano-HPLC-MS/MS and in silico analysis of complex protein mixtures. *J. Proteome Res.* 2004, *3*, 1138–1148.

[87] Kawakami, T., Tateishi, K., Yamano, Y., Ishikawa, T. *et al.*, Protein identification from product ion spectra of peptides validated by correlation between measured and predicted elution times in liquid chromatography/mass spectrometry. *Proteomics* 2005, *5*, 856–864.

[88] Varnum, S. M., Covington, C. C., Woodbury, R. L., Petritis, K. *et al.*, Proteomic characterization of nipple aspirate fluid: identification of potential biomarkers of breast cancer. *Breast Cancer Res. Treat.* 2003, *80*, 87–97.

[89] Strittmatter, E. F., Kangas, L. J., Petritis, K., Mottaz, H. M. *et al.*, Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. *J. Proteome Res.* 2004, *3*, 760–769.

[90] Qian, W. J., Liu, T., Monroe, M. E., Strittmatter, E. F. *et al.*, Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome. *J. Proteome Res.* 2005, *4*, 53–62.

[91] Nielsen, N.-P. V., Carstensen, J. M., Smedsgaard, J., Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatogr. A* 1998, *805*, 17–35.

[92] van Nederkassel, A. M., Daszykowski, M., Eilers, P. H., Heyden, Y. V., A comparison of three algorithms for chromatograms alignment. *J. Chromatogr. A* 2006, *1118*, 199–210.

[93] Callister, S. J., Dominguez, M. A., Nicora, C. D., Zeng, X. *et al.*, Application of the accurate mass and time tag approach to the proteome analysis of sub-cellular fractions obtained from *Rhodobacter* sphaeroides 2.4.1. Aerobic and photosynthetic cell cultures. *J. Proteome Res.* 2006, *5*, 1940–1947.

[94] Jennifer, S.D., Zimmer, M. E. M. W.-J. O. R. D. S., Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass Spectrom. Rev.* 2006, *25*, 450–482.

[95] Radulovic, D., Jelveh, S., Ryu, S., Hamilton, T. G. *et al.*, Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* 2004, *3*, 984–997.

[96] Listgarten, J., Neal, R. M., Roweis, S. T., Emili, A., Multiple alignment of continuous time series. MIT Press, Cambridge, MA 2005.

[97] Prakash, A., Mallick, P., Whiteaker, J., Zhang, H. *et al.*, Signal maps for mass spectrometry-based comparative proteomics. *Mol. Cell. Proteomics* 2006, *5*, 423–432.

[98] Bylund, D., Danielsson, R., Malmquist, G., Markides, K. E., Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data. *J. Chromatogr. A* 2002, *961*, 237–244.

[99] Stadalius, M. A., Gold, H. S., Snyder, L. R., Optimization model for the gradient elution separation of peptide mixtures by reversed-phase high-performance liquid chromatography : Verification of retention relationships. *J. Chromatogr. A* 1984, *296*, 31–59.

[100] Ishihama, Y., Method for detection of peptide sequence based on chromatography retention time. Patent: WO2007/013701 A1.

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[101] Ishihama, Y., Oda, Y., Tabata, T., Kawai, T. *et al.*, *Probability enhancement in protein identification by parent ion related parameters. 3rd Japan Human Proteome Organization Conference/JHUPO, Yokohama, Japan* 2005, P1–15.

[102] Shinoda, K., Yachie, N., Masuda, T., Sugiyama, N. *et al.*, HybGFS: a hybrid method for genome-fingerprint scanning. *BMC Bioinformatics* 2006, *7*, 479.

[103] Brunner, E., Ahrens, C. H., Mohanty, S.,Baetschmann, H. *et al.*, A high-quality catalog of the *Drosophila melanogaster* proteome. *Nat. Biotechnol.*2007, *25*, 576–583.

[104] Craig, R., Cortens, J. P., Beavis, R. C., The use of proteotypic peptide libraries for protein identification. *Rapid Commun. Mass Spectrom.* 2005, *19*, 1844–1850.

[105] Kuster, B., Schirle, M., Mallick, P., Aebersold, R., Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell. Biol.* 2005, *6*, 577–583.

[106] Ishihama, Y., Wei, F. Y., Aoshima, K., Sato, T. *et al.*, Enhancement of the efficiency of phosphoproteomic identification by removing phosphates after phosphopeptide enrichment. *J. Proteome Res.* 2007, *6*, 1139–1144.

[107] Steen, H., Jebanathirajah, J. A., Rush, J., Morrice, N., Kirschner, M. W., Phosphorylation analysis by mass spectrometry: myths, facts, and the consequences for qualitative and quantitative measurements. *Mol. Cell. Proteomics* 2006, *5*, 172–181.

[108] Kermani, B. G., Kozlov, I., Melnyk, P., Zhao, C. *et al.*, Using support vector machine regression to model the retention of peptides in immobilized metal-affinity chromatography. *Sens. Actuators B Chem.* 2007, *125*, 149–157.

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Transworld Research Network
37/661 (2), Fort P.O.
Trivandrum-695 023
Kerala, India

Clinical Application of Molecular Diagnosis in Cancer, Radiation Effect, and Human Diseases, 2009: ISBN: 978-81-7895-408-0 Editors: Eiso Hiyama and Keiko Hiyama

7. Diagnostic and prognostic molecular markers in breast cancer

Katsumasa Kuroi[1] and Masakazu Toi[2]

[1]Division of Clinical Trials and Research, Department of Surgery, Tokyo Metropolitan Cancer and Infectious Disease Center Komagome Hospital, Tokyo; [2]Department of Breast Surgery Graduate School of Medicine, Kyoto University, Kyoto, Japan

Abstract. This chapter aims to give a comprehensive insight into the possible clinical value of diagnostic and prognostic molecular markers in breast cancer. So far, genetic models for progression of breast cancer have not been developed; however, it is now well established that cancer is caused by the accumulation of genetic changes in a specific cell, and breast cancer can exhibit a tremendous range of alterations of oncogenes and tumor suppressor genes as well as allelic loss and microsatellite instability. Of clinical importance, *BRCA1* and *BRCA2* are the major breast cancer predisposition genes, and HER2 represents an excellent example of the translation of basic science to clinical practice. Moreover, molecular cloning of estrogen receptor (ER)β and ERβcx has lead to a paradigm shift in our understanding of estrogen action. In parallel, polymerase chain reaction (PCR) technology has brought the ability to amplify exponentially a previously undetectable amount of nucleic acid to a detectable level, providing a tool not only for molecular diagnosis, but also for molecular

Correspondence/Reprint request: Dr. Masakazu Toi, Department of Breast Surgery, Graduate School of Medicine Kyoto University, 54 Shogoin Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan
E-mail: toi@kuhp.kyoto-u.ac.jp

detection of micrometastasis or minimal residual disease. For molecular diagnosis, PCR assays for loss of heterozygosity, methylated alleles, or telomerase have now enabled noninvasive detection of small numbers of cancer cells. Moreover, reverse transcriptase–PCR could be used for the detection of micrometastases in lymph nodes, bone marrow, peripheral blood, and other body fluids. Candidate targets include carcinoembryonic antigen, cytokeratin 19, maspin, and mammaglobin. Thus, the rapid development of molecular technology has provided an opportunity for understanding the biology of breast cancer initiation and progression, and the heterogeneous nature of the disease. Ultimately, the use of these techniques will allow us to tailor the management of patients with breast cancer.

Introduction

In Japan, the incidence rate from female breast cancer has increased in recent years. In 2002, the age-adjusted incidence rate for female breast cancer was 52.2 per 100,000, ranking it the most frequent site of cancer in women [1]. So far, diagnostic and prognostic information has been based on cellular morphology, as little was known about the molecular pathology of breast cancer. Moreover, unlike colorectal cancer, genetic models for progression of breast cancer have not been developed. However, it is now well established that cancer is a disease of the genes, and that the phenotype of malignancy is often genetically determined. In agreement with this concept, knowledge has accumulated of genetic alterations in oncogenes and tumor suppressor genes as well as allelic loss and microsatellite instability, in breast cancer. Of clinical importance, several genes responsible for hereditary breast cancer have now been isolated, and an understanding of receptor function and co-regulatory molecules for the estrogen receptor (ER) and the progesterone receptor (PR) has begun to lead to better therapies. Further, targeting therapy toward molecular components preferentially overexpressed by breast cancer cells has become a widespread approach. In addition, recent advances in molecular technology have provided the tools not only for molecular diagnosis but also for molecular detection of micrometastases in lymph nodes (LNs), bone marrow (BM), peripheral blood (PB) and other body fluids. This chapter aims to give a comprehensive insight into the possible clinical value of molecular markers in breast cancer.

I. Gene alterations in breast cancer

Breast cancer can exhibit a tremendous range of genetic and chromosomal alterations. However, common lesions include oncogenes (*HER2*, *c-myc*,

cyclin D1, etc) and suppressor genes (*TP53, BRCA1, BRCA2, E-cadherin*, etc) (Table 1). In general, gene activation appears to be a common mechanism for oncogenes, whereas tumor suppressor genes are characterized primarily by point mutation, methylation, and loss of heterozygosity (LOH). Interestingly, a number of tumor suppressor genes involved in the initiation and progression of breast cancer have been mapped to chromosomes 17, 16, 11, 10 and 9, which have been reported to show a high rate of LOH in breast cancer. In the following section, we summarize the commonly described genetic alterations associated with hereditary and nonhereditary forms of breast cancer.

1. Genetic predisposition to breast cancer

Women who have close relatives with breast cancer are at an increased risk of developing breast cancer themselves. The majority of breast cancers are sporadic, and familial clustering of breast cancer may be coincidental; however, major breast cancer predisposition genes that are inherited in an autosomal dominant fashion may be responsible for an increased risk of breast cancer in some families (Table 2). Of all women with breast cancer, about 25 to 30% have a close family member with breast cancer, and 5 to 10% of breast cancer is due to cancer predisposition genes [2].

Cancer can occur in any cell, either somatic or germ line, that contains a nucleus, but inheritance requires a mutation in the germ line that will be passed on to the next generation at conception. Among several causative genes, mutation in *BRCA1* and *BRCA2* may be responsible for as much as 80 % of inherited breast cancer [3]. Both are tumor-suppressor genes, located on chromosomes 17q21 [4], and 13q12-13 [5] respectively, and their ubiquitously expressed protein products are involved in the maintenance of genome integrity, including DNA repair and recombination, checkpoint control of cell cycle, and transcription [6]. Despite few similarities between these products, they appear to co-operate in one or more DNA damage response pathways. So far, most of the mutations described in the *BRCA* genes are frameshift, nonsense, or splicing mutations that lead to premature protein truncation, and their discovery has made it possible to offer predictive genetic testing for women at high risk of breast cancer.

Germline mutations in *BRCA1* and *BRCA2* are associated with an increased risk for developing breast cancer and ovarian cancer, and to a lesser extent, colon cancer and prostate cancer for *BRCA1*, and male breast cancer and pancreas cancer for *BRCA2*. It is estimated that the cumulative risk of developing breast cancer among women with *BRCA1* mutations is 50% at age