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.

[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-

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-

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. Q. 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: W02007/013701 A1.

[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.

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

Table 1. Major oncogenes and tumor suppressor genes in breast cancer.

Genes	Locus	Products	Function
Ongogenes (amplified)			
c-myc	8q24	DNA binding protein	Cell cycle control, Differentiation, Adhesion, Apoptosis
HER2 (c-erbB-2)	17q21	HER2 oncoprotein	Growth factor receptor
Cycline D1 (CCDN1)	11q13	Cell cycle regulator (G1-S phase)	Cell cycle regulator, Inactivation of PB1
Suppressor genes (mutated or loss)			
RB1	13q14	pRB	Cell cycle regulator (S phase)
TP53	17p13.1	p53	Guardian of genome
BRCA1	17q21	BRCA1	Care taker gene to maintain genomic integrity during DNA replication
BRCA2	13q12-13	BRCA2	DNA repair
E-cadherin (CDH1)	16q24	E-cadherin	Cell adhesion
Nm23 (NME1 and 2)	17q21.3	Nucleoside diphosphate kinase A, B	Catalysis of phosphorylation of nucleoside diphosphates to nucleoside triphosphates
PTEN (MMAC1)	10q23	Phosphatidylinositol phosphatase	Negative regulation of cell adhesion, Migration
p21	6q21	Inhibitor of cyclin-dependent kinase, p21	Mediation of G1 arrest
p16 (CDKN2A)	9p21	p16	Inhibition of cyclin dependent kinase (CDK)

Table 2. Hereditary breast cancer syndromes.

Hereditary cancer syndrome	Genes	Clinical association
Familial breast and ovarian cancer	*BRCA1*	Breast cancer, Ovarian cancer, Colon cancer, Prostate cancer
Familial breast cancer	*BRCA2*	Breast cancer, Prostate cancer, Male breast cancer
Li-Fraumeri syndrome	*TP53*	Multiple cancers in childhood, Soft tissue sarcoma, Brain tumors, Leukemia Adrenocortical carcinomas, Bilateral breast cancer in women
Cowden disease	*PTEN*	Breast cancer, Gastrointestinal cancers, Thyroid diseases, Mucocutaneous lesions, Trichilemmomas Acral keratosis, Papillomatous papules, Mucosal lesions, Macroencephaly Lhermitte-Duclos disease
Ataxia telangiectasia	*ATM*	Breast cancer, Telangiectasia, Leukemia, Lymphoma, Occulocutaneous telangiectasia Cerebellar ataxia, Immune deficiency
Hereditary nonpolyposis colorectal cancer	*MMR*	Gastrointestinal cancer, Basal cell carcinoma of genitourinary tract, Endometrial cancer Keratocanthomas, Colonic diverticula, Breast cancer in women

50 and 87% at age 70; for women with *BRCA2* mutations, it ranges between 40% and 80% at age 70 [3, 7]. In addition, the lifetime risk of ovarian cancer in female carriers with *BRCA1* mutations is around 30 – 40%; in *BRCA2* mutation carriers, it is approximately 20%.

As these estimates are based on studies of families at high risk of breast cancer, often including as many as 10 affected members, the study design may lead to very high estimates of penetrance. In fact, several population-based studies have demonstrated that the frequency of *BRCA1* and *BRCA2* mutations is much lower than expected from data based on these rare and large high-risk families [8-10]. However, meta-analysis has revealed that women with such genetic abnormalities and a strong family history of breast cancer are still likely to possess a much higher risk for breast cancer than are women with such abnormalities but without a strong family history [11]. In Japanese women with a modest to minimal family risk, defined as those having at least one breast cancer or ovarian cancer patient in their first degree relatives, the frequencies of *BRCA1* and *BRCA2* mutations were 13.3% and 18.6%, respectively [8]. Families at high risk for carrying *BRCA1* and *BRCA2* mutations could be predicted by family history profiles characterized by first-degree relatives with ovarian cancer or breast cancer along with young age at diagnosis (40 years old or less), bilateral occurrence and increased number of affected relatives [8].

Thus, mutations of *BRCA1* confer increased risk for breast cancer, but it is not clear why the mutations are associated with breast cancer. One study of particular interest indicated that BRCA1 could inhibit ER transcription [12], suggesting that wild-type BRCA1 suppresses estrogen-dependent transcriptional pathways related to mammary epithelial cell proliferation and that loss of this ability contributes to tumorigenesis. In agreement with these findings, bilateral prophylactic oophorectomy is associated with a reduced breast cancer risk in women who carry a *BRCA1* mutation [13].

Clinically, BRCA-associated hereditary breast cancer displays unique clinicopathological features compared with sporadic breast cancer [14-18]. For example, there is a higher frequency of medullary carcinoma in *BRCA1* mutations and lobular carcinoma in *BRCA2* mutations. Furthermore, breast cancer in patients with *BRCA1* mutations generally presents with high grade, highly proliferative, aneuploid tumors. These features have also been observed in patients with *BRCA2* mutations, although this association appears weak. *BRCA1*-related tumors are generally receptor-negative, whereas this does not appear to be the case with *BRCA2*-related tumors. Similarly, a lower frequency of HER2 overexpression was observed in tumors with *BRCA1* mutations compared with sporadic cases [19, 20], whereas there is no difference in frequency between *BRCA2* related tumors and sporadic cases.

On the other hand, a higher frequency of *TP53* mutation has been observed in *BRCA*-related breast cancer [21], but this was not found in another report [20]. With respect to prognosis, the majority of studies have reported either no difference or only slightly worse survival in patients with *BRCA1* mutations [20, 22], despite the pathologic features of *BRCA1*-related tumors that are associated with poor prognosis. In contrast, *BRCA2*-related breast cancer seems more heterogeneous than the *BRCA1* phenotype, and not clearly different from sporadic forms [23].

2. Oncogenes

Among several oncogenes, the most important in breast cancer is *HER2/neu* (also known as *c-erbB-2*). This is located on 17q21, and encodes a transmembrane glycoprotein of 185 kilodaltons with intrinsic tyrosine kinase activity [24]. HER2 belongs to the epidermal growth factor (EGF) receptor (EGFR) family, which includes four known members (HER1, HER2, HER3, and HER4), and has molecular homology to EGFR. So far, no ligand has been identified for HER2; however, following ligand-dependent activation of HER1, HER3, or HER4 by EGF or heregulin, HER2 is activated by heterodimerization, resulting in activation of diverse subcellular signal transduction pathways [25-27], and thus HER2 appears to play a central role in HER signaling.

HER2 amplification is restricted to carcinoma of glandular epithelial origin, and is at least three times more common in breast cancer than in adenocarcinomas of other sites. Amplification or overexpression of HER2 has been found in 10-40% of primary breast cancer [25, 28]. Interestingly, HER2 has been identified in a high proportion (60-80%) of ductal carcinoma in situ of high grade, comedo-type, but is not common in the low nuclear grade forms [29, 30]. On the other hand, HER2 is very rarely expressed in lobular carcinoma in situ. HER2 has not been identified in benign proliferative disease including radial scar and atypical ductal hyperplasia [31]. However, low-level *HER2* amplification in benign breast biopsies was associated with increased risk of subsequent breast cancer [32]. In this study, HER2 overexpression was not detected in any of the benign breast tissues. Thus, HER2 may have a role in the inception of breast cancer.

Initially, *HER2* was identified as a marker of advanced-stage and poor prognosis, as it was associated with lymph nodes metastases, short relapse time, poor survival, and decreased response to endocrine therapy and chemotherapy [28, 33]. Because about one-third of ER and/or PR positive tumors do not respond to endocrine therapy, several retrospective studies have addressed the association of HER2 overexpression and decreased

efficacy of tamoxifen [34-36], but not all studies concur [37, 38]. Thus, the biologic classification of tumors by HER2 status as responsive or not to endocrine therapy has been difficult [39]. In this respect, however, it is interesting to note that tamoxifen can behave as an estrogen agonist in breast cancer cells expressing high levels of the co-activator, AIB1 and Her2, resulting in *de novo* tamoxifen resistance [40]. Moreover, in a study of preoperative endocrine therapy, ER positive, HER1 positive, and/or HER2 positive primary breast cancer responded well to letrozole, an aromatase inhibitor, but responded only infrequently to tamoxifen [41]. This result may be consistent with the observation that ER positive, HER2 positive tumors are highly estrogen dependent [42], and that MEKK1, a down-stream mediator of signaling by HER1 and HER2, activates the ER and stimulates the agonist activity of tamoxifen [43]. Aromatase inhibitors block the conversion of adrenal androgens to estrogens, which leaves the ER in target tissues deprived of ligand, and thus ER becomes inactive, monomeric and unable to bind DNA in ER positive, HER1 and/or HER2 positive tumors. As for chemotherapy, several studies have suggested that HER2 overexpression is predictive for the response to anthracycline-based adjuvant therapy [44-46]. Similarly, in one study using fluorescence in situ hybridization (FISH), anthracycline-based therapy was more effective than cyclophosphamide, methotrexate and 5-fluorouracil (CMF) in the subgroup of *HER2* amplified patients with node positive breast cancer [47]. In addition, the CALGB 8541 study found a significant dose-response effect of adjuvant cyclophosphamide, doxorubicin and 5-fluorouracil (CAF) in node-positive patients with amplified or overexpressed HER2 [44-46]. In this study, maximal benefit was achieved at 60 mg/m^2, which is now considered to be the standard dose of doxorubicin. Interestingly, in breast cancer cells, *HER2* amplification increased the activity of topoisomerase IIα (topo II), leading to increased sensitivity to topo II inhibitors such as anthracyclines [48]. The *topo II* gene is located on 17q12-q21, close to the *HER2* gene, and several studies have indicated that *topo II* amplification only occurs with concurrent *HER2* amplification [49-51]. Furthermore, co-amplification of *HER2* and *topo II* in breast cancer cells conferred extreme sensitive to anthracyclines, whereas *HER2* but not *topo II* amplified cells showed an intermediate degree of sensitivity to anthracyclines [49]. Thus, it is possible that the predictive value of HER2 regarding anthracycline-based therapy over ~~cyclophosphamide, methotrexate and 5-fluorouracil (~~CMF~~)~~ is explained by the concomitant amplification of topo II [47]. In contrast, in an anthracycline and taxane neoadjuvant chemotherapy study by Rody et al [52], HER-2, but not topo II, is highly predictive of tumor response. Based on these findings, anthracycline-based therapy seems to be preferred for HER2 positive

patients; however, when anthracyclines are contraindicated, CMF-like therapy should not be withheld from patients who overexpress HER2, as HER2 positivity does not convey absolute resistance to non-anthracycline-based regimens [39]. In contrast, in a recent study of anthracycline-based neoadjuvant chemotherapy in endocrine non-responsive Her2/neu-positive breast cancer [53], topo II gene status was predictive of pathological complete remission. Thus, further prospective studies should investigate the predictive value of co-amplification of *HER2* and *topo II*.

Importantly, *HER2* represents an excellent example of the translation of basic science to clinical practice. Its product, HER2, is a molecular target of the humanized monoclonal antibody trastuzumab (Herceptin) for the treatment of breast cancer [54, 55]. The activity of this agent is highly dependent on the HER2 status of the tumor. There are currently several different methods to evaluate HER2 status, the most common being immunohistochemical staining (IHC) and FISH [56]. Both results are often in agreement, but in cases with stain intensity of +1 or +2 by IHC, the interobserver agreement is poor and the predictive value is unsatisfactory for clinical use [57, 58]. Therefore, additional testing using FISH is recommended [58-60]. Similarly, equivocal result of FISH requires IHC assessment for final determination [56]. Otherwise, the proportion of phosphorylated HER2, which is indicative of activation, may provide more significant and additional information for the prediction of response to trastuzumab [61].

3. Tumor suppressor genes

In contrast to hereditary breast cancer, few mutations of *BRCA1* and *BRCA2* have been found in sporadic breast cancer [62, 63]. However, it is interesting to note that *BRCA1* mRNA levels were decreased in sporadic breast cancer compared with normal breast tissues [64]. Furthermore, BRCA1 expression was decreased or undetectable in high-grade breast cancer [65]. Thus, BRCA1 may also be involved in the progression of sporadic breast cancer. Aberrant methylation of the *BRCA1* promoter may be one mechanism for decreased expression of BRCA1 [66], and hypermethylation of the *BRCA1* promoter was strongly associated with negativity of ER and PR [67]. In contrast, a role for *BRCA2* inactivation remains unclear in sporadic breast cancer.

The *TP53* tumor suppressor gene is located on chromosome 17p13.1, and encodes a nuclear phosphoprotein. p53 is a key component of a cellular emergency response mechanism that converts a variety of intra- and extracellular stress signals such as DNA damage, as well as conflicts between

positive and negative proliferation stimuli, into growth arrest or apoptosis. *TP53*-dependent mechanisms eliminating abnormal cells are highly efficient and provide a potent shield against proliferation and survival of cells that have been exposed to damage or have suffered faults in cell duplication and interaction with their neighbors. Therefore, without damage to, or reduced activity of, the p53 pathway, the carcinogenesis process is virtually impossible. Indeed, *TP53* mutations, most of which are missense mutations, are the most common genetic alteration in human cancer. The mutated *TP53* is able to oligomerize with the product of the wild-type allele, thus exerting a dominant negative effect [68]. In addition, missense mutations in *TP53* increase the stability of p53 protein, resulting in a protein that can be detected by IHC. Therefore, IHC has been used extensively in the study of *TP53* alterations in human cancer, and 25-50% of breast tumors are positive for p53 by IHC. Clinically, *TP53* mutations or protein overexpression are frequently found in medullary and ductal carcinomas, but not in lobular, mucinous or papillary carcinomas, and breast cancers with *TP53* mutations are in general highly aggressive and associated with negative hormone receptor status and high histological grade [69]. Until now, a number of studies have linked TP53 overexpression to poor prognosis of breast cancer, and a meta-analysis has also suggested that *TP53* mutations confer an independent relative risk of 1.7 for both disease-free survival and overall survival [70]. However, IHC for TP53 will miss *TP53* deletions. Moreover, TP53 may have functional duality, as p53 reactivation in p53-deficient tumors could induce either programmed cell death or cellular senescence, butTP53 inhibitory therapies could also be used to prevent the recovery of therapy-damaged tumor cells [71]. Thus, there is still a lot to be discovered about p53.

4. Cell survival and apoptosis related genes

Telomeres are long tandem repeat sequences at the ends of all eukaryotic chromosomes, and function to maintain chromosomal integrity during cell division. As the majority of normal somatic cells do not express the ribonucleoprotein enzyme telomerase, they are unable to maintain their telomere length, and thus with each cell division the ends of chromosomes, consisting of the telomeric repeats TTAGGG, progressively erode. In contrast, it is now well established that telomerase activation is one of the mechanisms by which cancer cells evade death and achieve immortalization. Interestingly, use of the telomeric repeat amplification protocol (TRAP) assay has detected telomerase activity in more than 95% of advanced breast cancers, whereas it was absent from 19-32% of less advanced cancers [72]. In this study, some benign tumors of breast, such as fibroadenoma, also had low

but detectable telomerase activity. To date, the TRAP assay, which is based on the polymerase chain reaction (PCR), is the most widely used strategy for detecting telomerase activity; however, quantitative analysis of telomerase activity by methods such as fluorescent based real time PCR appears to be useful to distinguish cancer from benign tumors [73]. In addition, telomerase activity was detectable in ductal carcinoma in situ lesions [74]. These lesions appear to be heterogeneous with respect to telomerase activity, and telomerase activation may precede the development of invasive cancer. If so, telomerase activity may be a useful adjunct in stratifying the risk of developing invasive breast cancer in patients with ductal carcinoma in situ.

So far, very few data are available regarding the prognostic significance of telomerase activity in breast cancer, although one study has demonstrated that levels of telomerase activity correlated with clinical outcomes and several prognostic indicators [75]. In contrast, another study failed to show any prognostic significance of telomerase activity [76]. Regulation of telomerase activity correlates with the expression of two major components, human telomerase RNA (hTERC/hTR), and human telomerase reverse transcriptase (hTERT), and therefore detection of both hTR, either by Northern blot analysis or by *in situ* hybridization, and hTERT, by reverse transcriptase(RT)-PCR, appears to be an important subject for further studies in breast cancer [77, 78].

Apoptosis (programmed cell death) is a physiological process that is crucial to the growth and development of multi-cellular organisms, and abnormality in the control of apoptosis plays an important role in tumorigenesis. Among a large number of genes implicated in the regulation of apoptosis, *BCL2* was the first shown to lead to prolonged survival of cells by inhibiting apoptosis. BCL2 belongs to the BCL2 family, which comprises at least 17 pro-apoptotic and anti-apoptotic members. Some of these family members, such as BCL2, BAX and BCLX, are expressed in normal breast, and their altered expression in breast cancer has been investigated. For example, loss of BAX, a pro-apoptotic protein, is a prognostic indicator of poor response to chemotherapy and shorter survival in metastatic breast cancer [79]. In a similar study, BCL2 expression was positively correlated with improved survival following endocrine therapy or chemotherapy with CMF [80]. Moreover, BCL2 showed an inverse relationship with HER2 and p53, whereas a strong positive relationship was seen between BCL2 and ER [81]. The prognostic role of BCL2 was supported by a recent meta-analysis [82]. Thus, BCL2 positive and BCL2 negative breast cancers have distinct phenotypes, and BCL2 expression seems to be indicative of a less aggressive phenotype.

It is generally accepted that BCL2 mainly exerts its anti-apoptotic effect by heterodimerization with pro-apoptotic members of the BCL2 family such as BAX [83]. Interestingly, phosphorylation of BCL2 reduces heterodimer formation with BAX, decreasing or inactivating the antiapoptotic function of BCL2 [84, 85]. In agreement with these studies, Shitashige et al [86] demonstrated that the phosphorylation of BCL2 at serine 70 destroyed its binding affinity for BAX, and that BCL2 kinases or their activators were synthesized *de novo* after treatment with paclitaxel, leading to phosphorylation of BCL2 in a time- and dose-dependent manner. In this study, using polyclonal antibody against phosphorylated S70-BCL2 (pS70-Bcl2), the expression of pS70-Bcl2 correlated with the response to paclitaxel and docetaxel in breast cancer, suggesting that phosphorylation of BCL2 might be predictive for the response to taxanes.

Among the inhibitors of apoptosis (IAP), *survivin* is characterized by a unique structure with a single baculovirus IAP repeat and no zinc-binding domain (RING finger), and by selective tissue distribution in cancer but not in normal adjacent tissues [87]. Survivin is repressed in the G1 phase of the cell cycle, but is highly expressed in G2/M, and the expression of survivin is negatively regulated by p53 [88]. By binding the mitotic spindle during mitosis, survivin may counteract a default induction of apoptosis in G2/M phase [89, 90]. Thus, the expression of survivin is cell cycle-dependent, and the overexpression of survivin in cancer may overcome this apoptotic checkpoint and favor aberrant progression of transformed cells through mitosis. In fact, silencing survivin gene expression could promote apoptosis of human breast cancer cells through a caspase-independent pathway [91], and expression of survivin was positive in 70% of tumors and was strongly correlated with BCL2 expression and reduced apoptotic index in breast cancer [92].

5. Metastasis suppressor genes

E-cadherin is a transmembrane glycoprotein that functions as a cell adhesion molecule through its attachment to the actin cytoskeleton via α-, β -, γ-catenin. Therefore, loss of function of cadherin or catenin will promote metastatic progression. In fact, a high frequency of LOH at the *E-cadherin (CDH1)* locus on 16q24 has been reported in lobular carcinoma [93]. In contrast, although expression of E-cadherin was reduced in invasive breast cancer [94], inactivating point mutations in the *CDH1* gene have never been found in invasive ductal carcinoma [93, 95]. These findings may offer a molecular explanation for the typical scattered tumor cell growth in invasive lobular carcinoma. As for the clinical significance of E-cadherin expression,

it is interesting to note that a recent systematic review by Gould Rothberg et al [96] has demonstrated that reduced or absent E-cadherin expression significantly increase the risk of all-cause mortality.

Other important genes are *nm23-H1* (*NME1*) and *nm23-H2* (*NME2*) located on 17q21.3. These encode nucleoside diphosphate kinase A and B, respectively. Decreased expression is the key parameter that determines metastatic potential, and may occur through a variety of mechanisms, such as LOH or methylation [97, 98]. However, *NME1* is rarely inactivated by point mutations [99]. Reduced expression of *NME1* was associated with a highly metastatic phenotype in many epithelial tumors [100, 101], and high *NME1* mRNA levels were correlated with LN negative status, less aggressive disease and longer disease-free and overall survival in breast cancer [102]. However, association with LN metastasis and the long-term prognostic value of this gene has been questioned in other studies [103-106].

6. Multidrug resistance genes

Multidrug resistance (MDR) is one of the major reasons for the failure of cancer chemotherapy. MDR can be intrinsic or acquired, and is a multifactorial phenomenon involving several specific and nonspecific mechanisms. For example, increased level of thymidylate synthetase, altered expression of Topo II, and enhanced detoxification by glutathione-linked enzyme systems can cause clinical drug resistance [107-111]. Moreover, recent molecular investigations in MDR have identified several genes coding for multidrug resistance 1 (MDR1)/P-glycoprotein (P-gp), multidrug resistance-related protein 1 (MRP1), breast cancer resistant protein (BCRP), and lung resistance-related protein (LRP). These transmembrane proteins cause MDR either by decreasing the total intracellular retention of drugs or by redistributing intracellularly accumulated drugs away from target organelles [112].

The best characterized genes responsible for the MDR phenotype are *MDR1* and *MRP1*. Both belong to the ATP binding cassette (ABC) transporter gene family, and act as a membrane-associated ATP-dependent efflux pump or an ATP-dependent organic ion transporter, respectively [112]. Overexpression of MDR1 reduces the accumulation of chemotherapeutic drugs such as anthracyclines and taxanes inside cells, and results in the MDR phenotype. However, the contribution of these MDR transporters to chemotherapeutic failure in human breast cancer remains to be proven [113-116]. Several methods are available to investigate MDR. One approach to establishing the clinical significance of these transporters is to investigate the effects of specific inhibitors of each system. Furthermore, ^{99m}Tc-Sestamibi

scintigraphy appears to be a novel technique to characterize the function of MDR transporters *in vivo* [117, 118].

II. Hormone receptors

Estrogen exposure is the most important factor in breast tumorigenesis: risk for developing breast cancer increases with extended lifetime exposure to estrogens, whereas it decreases following oophorectomy. Furthermore, it has become clear that estrogen influences the expression of the PR through the ER, and that breast cancers are dependent for growth on estrogen or progesterone, acting through the ER or the PR, respectively. Both receptors belong to a superfamily of nuclear hormone receptors including other steroid hormone receptors, thyroid hormone receptors, vitamin D receptors and retinoic acid receptors. These receptors share a common structural and functional organization, and function as transcription factors when they are bound to their respective ligand. Fig. 1 shows the functional domains of the ER, designated A through F.

On estrogen binding at the ligand binding domain (LBD), ER forms a homodimer and binds to DNA through its DNA binding domain (DBD) with high affinity at a specific site, the estrogen response element (ERE), which exists in the enhancer region of target genes. Thus, the ER influences the expression of estrogen-responsive genes, including the PR, which are important in mitogenic signaling [119] (Fig. 2). These processes are regulated by receptor coregulating proteins. In the inactivated state, corepressor proteins bound to the receptor inhibit receptor function. When estrogen binds to the ER, corepressor proteins are replaced by coactivators which relax DNA and facilitate the recruitment of transcription factors to the complex, all of which serves to enhance transcription. Importantly, transcription is stimulated

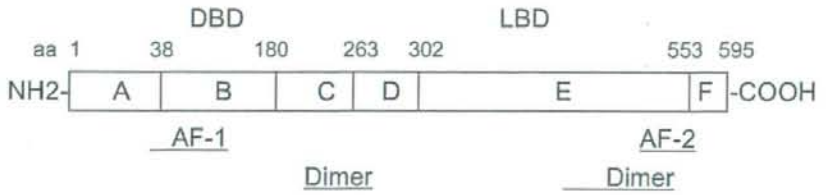

Figure 1. Schematic diagram of ERα. ER contains 595 amino acids with functional domains labeled A through F. DBD and LBD indicate DNA binding domain and ligand binding domain, respectively, and the regions important for dimerization (Dimer) and transactivation function (AF-1, AF-2) are shown.

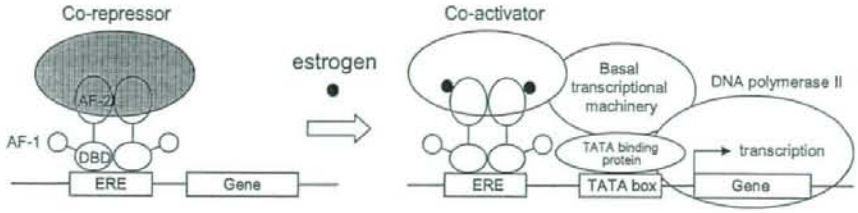

Figure 2. Model of agonist activity of ERα. In the inactivated state, DNA-bound histone deacetylase results in tight coiling of DNA, and co-repressor proteins bound to receptor inhibit receptor function. When ligand (estrogen) binds receptor, corepressor proteins are replaced by co-activators, and co-activators relax DNA and facilitate the recruitment of transcription factors to the complex, all of which serves to enhance transcription.

through at least two distinct transactivation domains, activation function (AF)-1, a hormone-independent amino-terminal ER activation domain located in the A/B region, and AF-2, a hormone-dependent, carboxy-terminal activation domain located in the E region. Although both AF-1 and AF-2 are required for maximal ER transcriptional activity in most cellular environments, AF-1and AF-2 can function independently [120]. When AF-2 is required for transcriptional activity, antiestrogens such as tamoxifen function as antagonists; however, where AF-2 is not required and AF-1 is sufficient for ER transcriptional activity, tamoxifen can function as a partial agonist on an ERE based reporter gene [121]. Thus, the function of AF-1 or AF-2 depends strongly upon promoter context, and the antagonist activity of tamoxifen arises from their intrinsic inability to activate AF-2. These observations provide a possible molecular explanation for the tissue-specific partial agonist properties of tamoxifen.

Clinically, approximately two-thirds of breast cancers express ER protein as assayed by ligand binding or IHC; however, IHC analysis of ER and PR is now becoming the predominant method [122]. Currently, ER and PR status are the only firmly established predictive factors for endocrine therapy. Endocrine therapy contributes significantly to prolongation of the disease-free survival, whereas breast cancers lacking ER rarely respond to endocrine therapy and are associated with a lower grade of histological differentiation, higher growth fraction and worse clinical outcome. Therefore, hormone resistance is a challenging problem in treatment of breast cancer, and loss of ER expression is an important cause of hormone resistance. The absence of PR from ER positive tumors generally implies that the ER is not active *in vivo*. In contrast, loss of ER may be due to a genomic deletion, mutations or rearrangement of the gene, or down-regulation of transcription at the protein