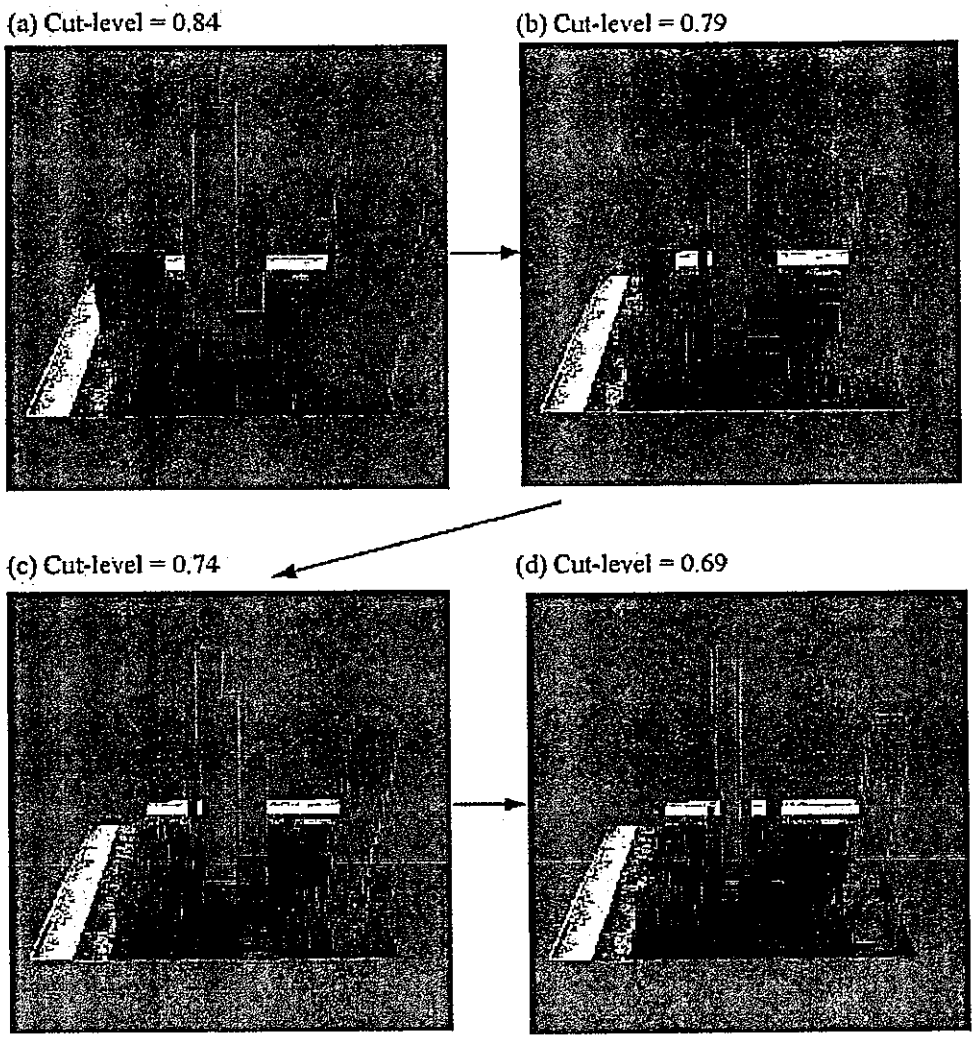

Figure 6. Interactive Changes of Cut-levels

In *CODM*, there is a risk that a small *overlap block* may be hidden in a large block. To avoid this problem, *CODM* allows the user to change the *cut level* interactively. If the user decreases the *cut level*, some small blocks that are hidden in larger blocks will emerge. By considering the homogeneity of clusters and the relationships with other gene information, the user can find important genes displayed as blocks in the *CODM*.

BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 4 2005, pages 439–444
doi:10.1093/bioinformatics/bti188

Multidimensional support vector machines for visualization of gene expression data

D. Komura [1,*], H. Nakamura[1], S. Tsutsumi[1], H. Aburatani[2] and S. Ihara[1]

[1]Research Center for Advanced Science and Technology and [2]Genome Science Division, Center for Collaborative Research, University of Tokyo, Tokyo 153-8904, Japan

Received on May 21, 2004; accepted on November 11, 2004
Advance Access publication December 17, 2004

ABSTRACT

Motivation: Since DNA microarray experiments provide us with huge amount of gene expression data, they should be analyzed with statistical methods to extract the meanings of experimental results. Some dimensionality reduction methods such as Principal Component Analysis (PCA) are used to roughly visualize the distribution of high dimensional gene expression data. However, in the case of binary classification of gene expression data, PCA does not utilize class information when choosing axes. Thus clearly separable data in the original space may not be so in the reduced space used in PCA.
Results: For visualization and class prediction of gene expression data, we have developed a new SVM-based method called multidimensional SVMs, that generate multiple orthogonal axes. This method projects high dimensional data into lower dimensional space to exhibit properties of the data clearly and to visualize a distribution of the data roughly. Furthermore, the multiple axes can be used for class prediction. The basic properties of conventional SVMs are retained in our method: solutions of mathematical programming are sparse, and nonlinear classification is implemented implicitly through the use of kernel functions. The application of our method to the experimentally obtained gene expression datasets for patients' samples indicates that our algorithm is efficient and useful for visualization and class prediction.
Contact: komura@hal.rcast.u-tokyo.ac.jp

1 INTRODUCTION

DNA microarray has been the key technology in modern biology and helped us to decipher the biological system because of its ability to monitor the expression levels of thousands of genes simultaneously. Since DNA microarray experiments provide us with huge amount of gene expression data, they should be analyzed with statistical methods to extract the meanings of experimental results.

A great number of supervised learning algorithms have been proposed and applied to classification of gene expression data (Golub *et al.*, 1999; Tibshirani *et al.*, 2002; Khan *et al.*, 2001). Support Vector Machines (SVMs) have been paid attention in recent years because of their good performance in various fields, especially in the area of bioinformatics including classification of gene expression data (Furey *et al.*, 2000). However, SVMs predict a class of test samples by projecting the data into one-dimensional space based on a decision function. As a result, information loss of the original data is enormous.

Some methods are used for projecting high dimensional data into lower dimensional space to clearly exhibit the properties of the data and to roughly visualize the distribution of the data. Principal Component Analysis (PCA) (Fukunaga, 1990) and its derivatives, e.g. Nonlinear PCA (Diamantaras and Kung, 1996) and Kernel PCA (Schölkopf *et al.*, 1998), are most widely used for this purpose (Huang *et al.*, 2003). One drawback of PCA analysis is, however, that class information is not utilized for class prediction because PCA chooses axes based on the variance of overall data. Thus clearly separable data in the original space may not be so in the reduced space used in PCA. Another method for visualization and reducing dimension of data is discriminant analysis. It chooses axes based on class information in terms of within- and between-class variance. However, it is reported that SVMs often outperform discriminant analysis (Brown *et al.*, 2000).

The main purpose of this paper is to cover the shortcoming of SVMs by introducing multiple orthogonal axes for reducing dimensions and visualization of gene expression data. To this end, we have developed multidimensional SVMs (MD-SVMs), a new SVM-based method that generates multiple orthogonal axes based on margin between two

*To whom correspondence should be addressed.

Komura *et al.* (2004) Multidimensional Support Vector Machines for Visualization of Gene Expression Data. Symposium on Applied Computing, Proceedings of the 2004 ACM symposium on Applied computing, 175-179; http://doi.acm.org/10.1145/967900.967936

classes to minimize generalization errors. The axes generated by this method reduce dimensions of original data to extract information useful in estimating the discriminability of two classes. This method fulfills the requirement of both visualization and class prediction. The basic properties of SVMs are retained in our method: solutions of mathematical programming are sparse, and nonlinear classification of data is implemented implicitly through the use of kernel functions.

This paper is organized as follows. In Section 2, we introduce the fundamental of SVMs. In Section 3, we describe the algorithm of MD-SVMs. In Section 4 and 5, we show numerical experiments on real gene expression datasets and reveal that our algorithm is effective for data visualization and class prediction.

1.1 Notation

$\mathbb{R}$ is defined as the set of real numbers. Each component of a vector $x \in \mathbb{R}^n, i = 1,\ldots,m$ will be denoted by $x_j, j = 1,\ldots,n$. The inner product of two vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ will be denoted by $x \cdot y$. For a vector $x \in \mathbb{R}^n$ and a scalar $a \in \mathbb{R}, a \leq x$ is defined as $a \leq x_i$ for all $i = 1,\ldots,n$. For an arbitrary variable x, x^k is just a name of the variable with upper suffix, not defined as k-th power of x.

2 SUPPORT VECTOR MACHINES

Since details of SVMs are fully described in the articles (Vapnik, 1998; Cristianini and Shawe-Taylor, 2000), we briefly introduce the fundamental principle of SVMs in this section. We consider a binary classification problem, where a linear decision function is employed to separate two classes of data based on m training samples $x_i \in \mathbb{R}^n, i = 1,\ldots,m$ with corresponding class values $y_i \in \{\pm 1\}, i = 1,\ldots,m$. SVMs map a data $x \in \mathbb{R}^n$ into a higher, probably infinite, dimensional space $\mathbb{R}^N$ than the original space with an appropriate nonlinear mapping $\phi : \mathbb{R}^n \to \mathbb{R}^N, n < N$. They generate the linear decision function of the form $f(x) = \mathrm{sign}(w \cdot \phi(x) + b)$ in the high dimensional space, where $w \in \mathbb{R}^N$ is a weight vector which defines a direction perpendicular to the hyperplane of the decision function, while $b \in \mathbb{R}$ is a bias which moves the hyperplane parallel to itself. The optimal decision function given by SVMs is a solution of an optimization problem

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i,$$

$$\text{s.t. } y_i(w \cdot \phi(x_i) + b) \geq 1 - \xi_i, \quad i = 1,\ldots,m, \xi \geq 0, \quad (1)$$

with $C > 0$. Here, $\xi \in \mathbb{R}^m$ is a vector whose elements are slack variables and $C \in \mathbb{R}$ is a regularization parameter for penalizing training errors. When $C \to \infty$, no training errors are allowed, and thus this is called hard margin classification. When $0 < C < \infty$, this is called soft margin classification because it allows some training errors. Note that a geometric margin γ between two classes is defined as $\frac{1}{\|w\|^2}$. The optimization problem formalizes the tradeoff between maximizing margin and minimizing training errors. The problem is transformed into its corresponding dual problem by introducing lagrange multiplier $\alpha \in \mathbb{R}^m$ and replacing $\phi(x_i) \cdot \phi(x_j)$ by kernel function $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ to be solved in an elegant way of dealing with a high dimensional vector space. The dual problem is

$$\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i, x_j) + \sum_{i=1}^{m}\alpha_i,$$

$$\text{s.t. } 0 \leq \alpha \leq C, \sum_{i=1}^{m}\alpha_i y_i = 0. \quad (2)$$

By virtue of the kernel function, the value of the inner product $\phi(x_i) \cdot \phi(x_j)$ can be obtained without explicit calculation of $\phi(x_i)$ and $\phi(x_j)$. Finally, the decision function becomes $f(x) = \mathrm{sign}\left(\sum_{i=1}^{m}\alpha_i y_i K(x_i, x) + b\right)$. by using kernel functions between training samples $x_i, i = 1,\ldots,m$ and a test sample x.

3 MULTIDIMENSIONAL SUPPORT VECTOR MACHINES

In order to overcome the drawback that SVMs cannot generate more than one decision function, we propose a SVM-based method that can be used for both data visualization and class prediction in this section. We call this method multi-dimensional SVMs (MD-SVMs). We deal with the same problem as mentioned in Section 2. Conventional SVMs give an optimal solution set (w, b, ξ) which corresponds to a decision function, while our MD-SVMs give the multiple sets $(w^k, b^k, \xi^k), k = 1,2,\ldots,l$ with $l \leq n$, so that all the directions w_k are orthogonal to one another. The orthogonal axes can be used for reducing the dimension of original data and data visualization in three dimensional space by means of projection. Here the first set (w^1, b^1, ξ^1) is equivalent to that obtained by conventional SVMs. Now we only refer to the steps of obtaining $(w^k, b^k, \xi^k), k = 2,3,\ldots,l$. In practice, the k-th set $(w^k, b^k, \xi^k) k = 2,3,\ldots,l$ are found with iterative computations of the optimization problem

$$\min_{w^k,\xi^k} \frac{1}{2}\|w^k\|^2 + C\sum_{i=1}^{m}\xi_i^k,$$

$$\text{s.t. } y_i(w^k \cdot \phi(x_i) + b^k) \geq 1 - \xi_i^k, i = 1,\ldots,m,$$

$$\xi^k \geq 0, w^k \cdot w^j = 0, j = 1,\ldots,k-1. \quad (3)$$

This problem differs from that of conventional SVMs in the last constraint $w^k \cdot w^j = 0$. The weight vector $w^j, j = 1,\ldots,k-1$ should be computed in advance by solving

other optimization problems (3). The optimization problem is modified by introducing lagrange multipliers $\alpha^k, \gamma^k \in \mathbb{R}^m$, $\beta^k \in \mathbb{R}^{k-1}$ and kernel functions. The primal Lagrangian is

$$L(w^k, b^k, \xi^k) = \frac{1}{2}\|w^k\|^2 + C\sum_{i=1}^{m}\xi_i^k + \sum_{i=1}^{m}\alpha_i^k(1 - \xi_i^k - y_i(w^k \cdot \phi(x_i) + b^k)) + \sum_{j=1}^{k-1}\beta_j^k(w^k \cdot w^j) - \sum_{i=1}^{m}\gamma_i^k\xi_i. \quad (4)$$

Consequently, the optimization problem is

$$\max_{\alpha^k,\beta^k} -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i^k\alpha_j^k y_i y_j K(x_i, x_j) + \frac{1}{2}\sum_{i=1}^{k-1}\beta_i^k\beta_i^k(w^i \cdot w^i) + \sum_{i=1}^{m}\alpha_i^k,$$

$$\text{s.t. } 0 \le \alpha^k \le C, \sum_{i=1}^{m}\alpha_i^k y_i = 0,$$

$$\sum_{i=1}^{m}\alpha_i^k y_i(\phi(x_i) \cdot w^j) = 0, j = 1, \ldots, k-1 \quad (5)$$

Here $\phi(x_p) \cdot w^q$ and $w^p \cdot w^p$ are calculated recursively as follows:

$$\phi(x_p) \cdot w^q = \sum_{i=1}^{m}\alpha_i^q y_i K(x_p, x_i) - \sum_{i=1}^{q-1}\beta_i^q(\phi(x_p) \cdot w^i), \quad (6)$$

$$w^p \cdot w^p = \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i^p\alpha_j^p y_i y_j K(x_i, x_j) - \sum_{i=1}^{m}\sum_{j=1}^{p-1}\alpha_i^p y_i\beta_j^p(\phi(x_i) \cdot w^j) + \sum_{i=1}^{p-1}\beta_i^p\beta_i^p(w^i \cdot w^i) - \sum_{i=1}^{m}\sum_{j=1}^{p-1}\alpha_i^p y_i\beta_j^p(\phi(x_i) \cdot w^j), \quad (7)$$

where $\phi(x_p) \cdot w^1 = \sum_{i=1}^{m}\alpha_i^1 y_i K(x_p, x_i)$ and $w^1.w^1 = \sum_{i=1}^{m}\alpha_i^1 y_i(\phi(x_i), w^1)$. As can be seen, there is no need to calculate nonlinear map of data $\phi(x)$ in problem (5) because all nonlinear mappings can be replaced with kernel functions.

Note that this optimization problem is a nonconvex quadratic problem when k is more than 1. As a consequence, the optimal solutions are not easy to be obtained. In Section 4, we use local optimum for numerical experiments when k is 2 or 3. We note the experimental results are still encouraging.

The corresponding Karush–Kuhn–Tucker conditions are

$$\alpha_i^k\{1 - \xi_i^k - y_i(w^k \cdot \phi(x_i) + b^k)\} = 0, \quad (8)$$

$$\xi_i^k(\alpha_i^k - C) = 0, i = 1, \ldots, m. \quad (9)$$

These are exactly the same as conventional SVMs. We highlight the other properties conserved from conventional SVMs:

- Projecting data into high dimensional space is implicit, using kernel functions to replace inner products.
- The solutions α^k of the optimization problem is sparse. Then the corresponding decision function depends only on few 'Support Vectors'.

Since each decision function is normalized independently to hold $w^k \cdot \phi(x_i) + b^k = y_i$ for $i = 1, \ldots, m$, data scales of the axes should be aligned with first axis ($k = 1$) for visualization. The margin γ^k, the L2-distance between support vectors of each class of k-th axis, is

$$\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i^k\alpha_j^k y_i y_j K(x_i, x_j) - \sum_{i=1}^{k-1}\beta_i^k\beta_i^k(w^i \cdot w^i)\right)^{-\frac{1}{2}}. \quad (10)$$

So a scaling factor $s^k = \gamma^1/\gamma^k$ is

$$\sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i^1\alpha_j^1 y_i y_j K(x_i, x_j)}{\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i^k\alpha_j^k y_i y_j K(x_i, x_j) - \sum_{i=1}^{k-1}\beta_i^k\beta_i^k(w^i \cdot w^i)}}. \quad (11)$$

The decision function of k-th step has the form $f^k(x) = \text{sign}\left(\sum_{i=1}^{m}\alpha_i^k y_i K(x_i, x) + b^k\right)$. Since the right hand side of the equation has the function of projecting original data into one dimensional space, the data can be plot in up to three dimensional space for visualization. The coordinate of data $x \in \mathbb{R}^m$ in three dimensional space is

$$(s^{k_1}g^{k_1}(x), s^{k_2}g^{k_2}(x), s^{k_3}g^{k_3}(x)), \quad (12)$$

where $g^k(x) = \sum_{i=1}^{m}\alpha_i^k y_i K(x_i, x) + b^k$. The space represents a distribution of data clearly based on the margin between two classes.

4 NUMERICAL EXPERIMENTS

4.1 Method

In order to confirm the effectiveness of our algorithm, we have performed numerical experiments. MD-SVMs can generate multiple axes, up to the number of features. Here we choose three axes, $k = 1, 2, 3$, to simplify the experiments. When k is

2 or 3, we use local optimum in problem (5) since it is difficult to obtain the global solutions. In our experiments, we carry out hold-out validation because cross-validation changes decision functions every time the dataset is split. Then we compare the results obtained by MD-SVMs with those obtained by PCA.

In the experiments, the expression values for each of the genes are normalized such that the distribution over the samples has a zero mean and unit variance. Before normalization, we discard genes in the dataset with the overall average value less than 0.35. Then we calculate a score $F(x(j)) = |(\mu^+(j)-\mu^-(j))/(\sigma^+(j)+\sigma^-(j))|$, for the remaining genes. Here $\mu^+(j)(\mu^-(j))$ and $\sigma^+(j)(\sigma^-(j))$ denote the mean and standard deviation of the j-th gene of the samples labeled $+1(-1)$, respectively. This score becomes the highest when the corresponding expression levels of the gene differ most in the two classes and have small deviations in each class. We select 100 genes with the highest scores and use them for hold-out validation. These procedures for gene selection are done only for training data for fair experiments.

The regularization parameter C in problem (5) is set to 1000. This value is rather large but finite because we would like to avoid ill-posed problems in a hard margin classification. We choose linear kernel $K(x_i, x_j) = x_i \cdot x_j$ and RBF kernel $K(x_i, x_j) = \exp -\gamma \| x_i - x_j \|^2$ with $\gamma = 0.001$ in the experiments of MD-SVMs.

4.2 Materials

Leukemia dataset (Golub et al., 1999) This gene expression dataset consists of 72 leukemia samples, including 25 acute myeloid leukemia (AML) samples and 47 acute lymphoblastic leukemia (ALL) samples. They are obtained by hybridization on the Affymetrix GeneChip containing probe sets for 7070 genes. Training set contains 20 AML samples and 42 ALL samples. Test set contains 5 AML samples and 5 ALL samples. AML samples are labeled +1 and ALL samples are labeled −1.

Lung tissue dataset (Bhattacharjee et al., 2001) This dataset consists of 203 samples from lung tissue, including 16 samples from normal tissue and 187 samples from cancerous tissue, and is obtained by hybridization on the Affymetrix U95A Genechip containing probe sets for 12558 genes. Training set includes 13 samples from normal tissue and 157 samples from cancerous tissue. Test set includes 3 samples from normal tissue and 30 samples from cancerous tissue. Samples from normal tissue are labeled +1 and samples from cancerous tissue are labeled −1.

5 RESULTS AND DISCUSSION

The results of numerical experiments are shown in Figure 1, and Tables 1 and 2. The distributions obtained by MD-SVMs on the leukemia dataset and the lung tissues dataset are given in Figure 1-(1) and 1-(3), respectively. Those obtained by PCA are given in Figure 1-(2) and 1-(4), respectively. The number of misclassified samples by MD-SVMs are summarized in Table 1 and 2. In these tables, the class of the samples is predicted based on decision functions $f^k(x), k = 1,2,3$, corresponding to each of the three axes.

Figure 1-(1) and 1-(3) illustrate that MD-SVMs are likely to separate the samples of each class in all the three directions. However, as shown in Figure 1-(2) and 1-(4), PCA does not separate the samples in the directions of the 2nd or the 3rd axis. These axes by PCA are dispensable with the objective of visualization for class prediction. In other words, MD-SVMs gather the plots of the samples into the appropriate clusters of each class, while PCA rather scatters them. Furthermore, in the distribution by MD-SVMs for the lung tissues dataset, one sample outlies from correct clusters (indicated by arrows in Figure 1-(3)). Though this sample also seems to be an outlier in the distribution by PCA (also indicated in Figure 1-(4)), the outlier significantly deviates in MD-SVMs. This may arise from the fact that MD-SVMs can separate the samples in all the directions. These observations indicate that MD-SVMs are well suited for visualizing in binary classification problems.

The significant advantage of MD-SVMs over PCA is the ability to predict the classes. MD-SVMs can predict the classes of samples based on the decision functions $f^k(x)$ without extra computation, while PCA cannot. The predicted class of a sample should be matched by the all the decision functions in an ideal case. However that does not always occur as seen in Tables 1 and 2. In such cases, the simplest method for prediction is to use only the 1st axis, which corresponds to the decision function generated by conventional SVMs. The idea is supported by the fact that the 1st decision function classifies the samples most correctly in almost all cases in Tables 1 and 2. The more advanced method is weighted voting. Scaling factor or normalized objective values in problem (5) are the candidate of the weight.

Multiple decision functions generated by MD-SVMs are useful for outlier detection. Samples misclassified by multiple decision functions may be mis-labeled or categorized into unknown classes. For example, see the column '3 axes' of test sample of the lung tissues dataset with RBF kernel in Table 2. This sample is misclassified by all decision functions, so we can say that this data contains some experimental error. The hierarchical clustering method also supports our result. These results indicate that MD-SVMs can be used for finding candidates of outliers.

6 CONCLUSION

For both visualization and class prediction of gene expression data, we propose a new method called Multidimensional Support Vector Machines. We formulate the method as a quadratic program and implement the algorithm. This is motivated by the following facts: (1) SVMs perform better than the other classification algorithms, but they generate only one axis for class prediction. (2) PCA chooses multiple

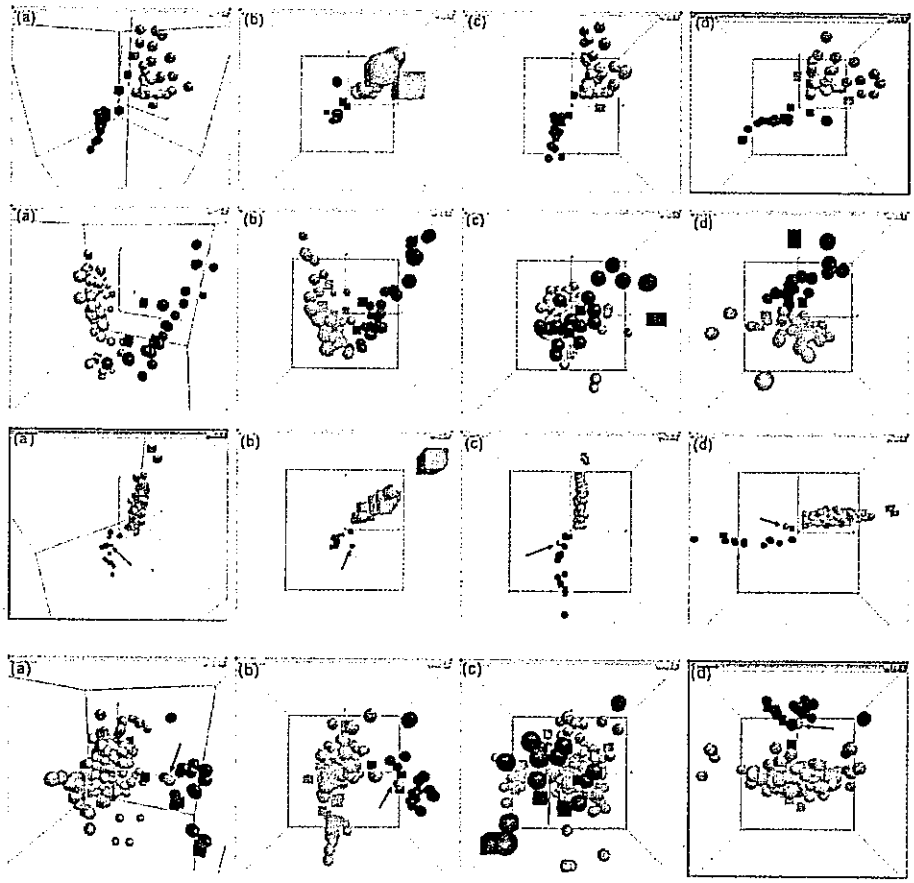

Fig. 1. (Top row) Distribution obtained by MD-SVMs for the leukemia dataset with linear kernel. (Second row) Distribution obtained by PCA on the leukemia dataset. (Third row) Distribution obtained by MD-SVMs for the lung tissues dataset with linear kernel. The sample indicated by arrows appears to be an outlier. (Fourth row) Distribution obtained by PCA for the lung tissues dataset. The sample indicated by arrows is the same as in the third row but with less deviates. (a) Cross shot, (b) 1st axis (x axis) and 2nd axis (y axis), (c) 2nd axis (x axis) and 3rd axis (y axis), (d) 3rd axis (x axis) and 1st axis (y axis). Black objects and white objects indicate AML samples (or normal tissues) ALL samples (or cancreous tissues), respectively. Training data and test data are expressed as a sphere and a cube, respectively.

Table 1. Number of classification errors in the MD-SVMs for the leukemia dataset. The columns 'n-th axis', $n = 1, 2, 3$, indicates the number of samples misclassified by n-th decision function. The columns 'n axes', $n = 1, 2, 3$, indicates the number of samples misclassified by n decision functions

Kernel	Sample	# of samples	1st axis	2nd axis	3rd axis	1 axis	2 axes	3 axes
Linear	Training	62	0	1	2	1	1	0
RBF	Training	62	0	2	7	5	2	0
Linear	Test	10	1	1	2	2	1	0
RBF	Test	10	0	2	0	2	0	0

Table 2. Number of classification errors in the MD-SVMs on the lung dataset. See the caption of Table 1 for other explanation

Kernel	Sample	# of samples	1st axis	2nd axis	3rd axis	1 axis	2 axes	3 axes
Linear	Training	170	0	1	1	0	1	0
RBF	Training	170	0	3	5	2	3	0
Linear	Test	33	1	0	0	1	0	0
RBF	Test	33	1	1	1	0	0	1

orthogonal axes, but it cannot predict classes of samples without other classification algorithms. We have tried to cover the shortcomings of both methods. MD-SVMs choose multiple orthogonal axes, which correspond to decision functions, from high dimensional space based on a margin between two classes. These multiple axes can be used for both visualization and class prediction.

Numerical experiments on real gene expression data indicate the effectiveness of MD-SVMs. All axes generated by MD-SVMs are taken into account for separating class of samples, while the 2nd and the 3rd axes by PCA are not. The samples in the distributions by MD-SVMs gather into appropriate clusters more vividly than those by PCA. MD-SVMs can predict the classes of the samples with multiple decision functions. We also indicate that MD-SVMs are useful for outlier detection with multiple decision functions.

There are several future works to be done on MD-SVMs: (1) application of our method to wider variety of gene expression datasets, (2) investigation of gene selection for preprocess of analysis and (3) investigation on class prediction method with multiple decision functions. Firstly, the use of more suitable samples may show that the axes chosen by MD-SVMs separate samples more clearly than those by PCA. Secondly, since the conventional SVMs show good generalization performance especially with large number of features, it is expected that MD-SVMs show much better performance than PCA with increasing the number of genes used in the numerical experiments. Since the element of weight vector generated by SVMs is one of the measures of discrimination power of the corresponding genes (Guyon *et al.*, 2002), that generated by MD-SVMs can be used for gene selection. Thirdly, the classification with probability as well as the weighted voting mentioned in Section 4 may be achieved in our scheme since the conventional SVMs have been already expanded for the purpose with sigmoid functions (Platt, 1999). We hope that our method sheds some lights on the future study of gene expression experiments.

REFERENCES

Bhattacharjee,A., Richards,W., Staunton,J., Li,C., Monti,S., Vasa,P., Ladd,C., Beheshti,J., Bueno,R., Gillette,M. *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.

Brown,M., Grundy,W., Lin,D., Cristianini,N., Sugnet,C., Furey,T., Ares,M. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, NY.

Diamantaras,K. and Kung,S. (1996) *Principal Component Neural Networks Theory and Applications*. John Wiley & Sons, NY.

Fukunaga,K. (1990) *Introduction to Statistical Pattern Recognition*. Academic Press, NY.

Furey,T., Cristianini,N., Duffy,N., Bednarski,D., Schummer,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Golub,T., Slonim,D., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J., Coller,H., Loh,M., Downing,J., Caligiuri,M., Bloomfield,C. and Lander,E. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Guyon,I., Weston,J., Barnhill,S. and Vapnik,V. (2002) Gene selection for cancer classification using support vector machines. *J. Machine Learn.*, **46**, 389–422.

Huang,E., Ishida,S., Pittman,J., Dressman,H., Bild,A., Kloos,M., D'Amico,M., Pestell,R., West,M. and Nevins,J. (2003) Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat. Genet.*, **34**, 226–230.

Khan,J., Wei,J., Ringnér,M., Saal,L., Ladanyi,M., Westermann,F., Berthold,F., Schwab,M., Antonescu,C., Peterson,C. and Meltzer,P. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.

Platt,J. (1999) *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*. MIT Press, Cambridge, MA.

Schölkopf,B., Smola,A. and Müller,K. (1998) Non-linear component analysis as a kernel eigenvalue problem. *Neural Comput.*, **10**, 1299–1319.

Tibshirani,R., Hastie,T., Narasimhan,B. and Chu,G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Vapnik,V. (1998) *Statistical Learning Theory*. John Wiley & Sons, NY.

Ab initio Quantum Mechanical Study of the Binding Energies of Human Estrogen Receptor α with Its Ligands: An Application of Fragment Molecular Orbital Method

KAORI FUKUZAWA,[1] KAZUO KITAURA,[2] MASAMI UEBAYASI,[3] KOTOKO NAKATA,[4] TSUGUCHIKA KAMINUMA,[5] TATSUYA NAKANO[4]

[1]*Biotechnology, Science Solutions, Mizuho Information & Research Institute, Inc., 2-3 Kanda Nishiki-cho, Chiyoda-ku, Tokyo 101-8443, Japan*

[2]*Research Institute for Computational Sciences, National Institute of Advanced Industrial Science and Technology, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan*

[3]*Institute for Biological Resources and Functions, National Institute of Advanced Industrial Science and Technology, 1-1-1 Higashi, Tsukuba, Ibaraki 305-8566, Japan*

[4]*Division of Safety Information on Drug, Food and Chemicals, National Institute of Health Sciences, 1-18-1 Kamiyoga, Setagaya-ku, Tokyo 158-8501, Japan*

[5]*Center for Quantum Life Science, Hiroshima University, 1-3-1 Kagamiyama, Higashihiroshima-shi, Hiroshima, 739-850, Japan*

Received 28 December 2003; Accepted 10 August 2004
DOI 10.1002/jcc.20130
Published online in Wiley InterScience (www.interscience.wiley.com).

Abstract: We have theoretically examined the relative binding affinities (RBA) of typical ligands, 17β-estradiol (EST), 17α-estradiol (ESTA), genistein (GEN), raloxifene (RAL), 4-hydroxytamoxifen (OHT), tamoxifen (TAM), clomifene (CLO), 4-hydroxyclomifene (OHC), diethylstilbestrol (DES), bisphenol A (BISA), and bisphenol F (BISF), to the α-subtype of the human estrogen receptor ligand-binding domain (hERα LBD), by calculating their binding energies. The *ab initio* fragment molecular orbital (FMO) method, which we have recently proposed for the calculations of macromolecules such as proteins, was applied at the HF/STO-3G level. The receptor protein was primarily modeled by 50 amino acid residues surrounding the ligand. The number of atoms in these model complexes is about 850, including hydrogen atoms. For the complexes with EST, RAL, OHT, and DES, the binding energies were calculated again with the entire ERαLBD consisting of 241 residues or about 4000 atoms. No significant difference was found in the calculated binding energies between the model and the real protein complexes. This indicates that the binding between the protein and its ligands is well characterized by the model protein with the 50 residues. The calculated binding energies relative to EST were very well correlated with the experimental RBA (the correlation coefficient $r = 0.837$) for the ligands studied in this work. We also found that the charge transfer between ER and ligands is significant on ER–ligand binding. To our knowledge, this is the first achievement of *ab initio* quantum mechanical calculations of large molecules such as the entire ERαLBD protein.

 J Comput Chem 26: 1–10, 2005

Key words: ***ab initio* fragment molecular orbital (FMO) method; estrogen receptor α; ligand-binding domain; binding energy; charge transfer**

Introduction

The steroid hormone estrogens play important roles in the regulation of growth, differentiation, and homeostasis in a variety of tissues. These effects are induced by the binding of estrogens to the intranuclear estrogen receptors (ER), for which two subtypes, ERα and ERβ, have been identified.[1–5] ERs are members of the nuclear receptor (NR) superfamily that includes other endocrine receptors such as the glucocorticoid receptor (GR), androgen receptor (AR), retinoic acid receptor (RAR), and so-called orphan receptors. These NRs function as ligand-activated transcriptional factors. The NRs with ligands and their coactivator form complexes that bind to

Correspondence to: K. Fukuzawa; e-mail: Kaori.fukuzawa@gene.mizuho-ir.co.jp

specific sequences called response elements in the promoter regions of the target genes, and then transcription of various target genes is mediated. The NR family possesses common domain functions and structures—that is, variable N-terminal transactivation domains, conserved DNA-binding domains (DBD), variable hinge regions, conserved ligand-binding domains (LBD), and variable C-terminal regions.[6-11] The characteristic ligand-induced motion of NRs is the conformational change of the helix12 (H12) at the C-terminal of LBD. When agonists bind to NR LBDs, the active conformation of H12 is formed, which possesses the coactivator binding surface. Antithetically, when antagonists bind to LBD, the coactivator binding surface is not produced because H12 is prevented from reaching its correct position.[12,13]

Because a variety of unknown compounds might bind to ER LBD and exert hormone-like effects on humans and wildlife, ER has become an important research target for the development of therapeutic agents[5,14,15] as well as for the screening of endocrine disruptors.[16,17] A number of experimental and theoretical studies have been carried out to clarify the mechanism of the ER–ligand binding. The binding affinities of several compounds to ER have been measured experimentally, and some of them are shown in Table 1,[4] as the relative binding affinities (RBAs) of each compound relative to the xenoestrogen, 17β-estradiol (EST, 1). Since 1997, the crystal structures of ER LBD with several ligands have been solved including 1ERE, 1ERR, 3ERT, and 3ERD, complexed with 17β-estradiol, raloxifen, 4-hydroxytamoxifen, and diethylstilbestrol, respectively.[12,13,18-21] These crystal structures have revealed the mode of binding between ERs and the ligands in detail. Figure 1 shows the ligand-binding site of hERα with 17β-estradiol, which is constituted by the ligand and the surrounding polar and charged amino acid residues.[12,18] The hydrogen bonds, which occur directly between the ligand and the surrounding residues, as well as through the mediation of a single water molecule, have been shown to stabilize the ER–ligand binding.[12,13,18]

Using available three-dimensional structures on a public database, numerous computational studies have been performed, most of them using empirical force fields.[22-28] However, force field approaches may not be reliable enough to predict binding energies between proteins and ligands. On the other hand, *ab initio* quantum mechanical calculations have played an important role in the study

Table 1. RBA for ERα and ERβ Subtypes [Experimental Values for Humans,[4] RBA of Each Ligand Is Calculated as Ratio of Concentrations of EST or Competitor Required to Reduce the Specific Radioligand Binding by 50% (=Ratio of IC_{50} Values)].

Ligand	ERα	ERβ
EST (1)	100[a]	100
DES (2)	236[a]	221
RAL (3)	69[a]	16[a]
OHT (4)	257[a]	232
GEN (5)	4	87[a]
TAM (6)	4	3
ESTA (9)	7	2
BISA (10)	0.01	0.01

[a]PDB crystal structures are published.

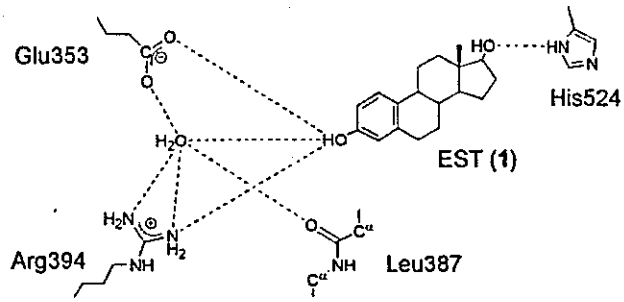

Figure 1. Hydrogen bond network at the ligand binding site of ER complexed with EST (1). Dotted lines indicate hydrogen bonds.

of molecular interactions as well as of the structures and properties of molecules. Although there has been increasing interest in the application of quantum mechanical calculations to bio-macromolecules, the requirement of huge computational resources has made such application very difficult. Recently, several research groups have worked on this issue and have succeeded in the quantum mechanical calculation of bio-macromolecules such as cytochrome *c* with about 100 amino acid residues.[29-34] Hoping to establish a time-saving and versatile computational procedure for bio-macromolecules, we are developing a quantum mechanical method and a program package—that is, the fragment molecular orbital (FMO) method[35-39] and the ABINIT-MP program (available from http://www.fsis.iis.u-tokyo.ac.jp/en/result/software/). The FMO method enables the calculation of proteins consisting of ~500 amino acid residues and polynucleotides of similar size with *ab initio* MO quality and with practical computational time.

In the present article, we report the first systematic study of the *ab initio* quantum mechanical calculations of proteins with more than 200 amino acid residues: the FMO method at the HF/STO-3G level was applied to the calculations of the binding energies of ligands with hERα LBD (241 residues).[40]

Molecular Modeling

The ligand molecules examined here are displayed in Figure 2, with the binding sites indicated in red and blue: the xenoestrogens EST (1) and ESTA (9), the phytoestrogen GEN (5), the synthetic estrogens DES (2), RAL (3), OHT (4), TAM (6), OHC (7), and CLO (8), and the industrial chemicals BISA (10) and BISF (11). In this article, we focused on the hydrogen bond networks in Figure 1, and performed the calculations in two steps. First, the most stable geometries of the hydrogen bond network were calculated using small model complexes, MODEL3, as described below. Second, the binding energies were calculated using larger model complexes, MODEL1 and MODEL2, with the hydrogen-bond geometries obtained at the first step.

Model Systems

We used three models for the ER–ligand complexes. Figure 3 shows MODEL 1 and MODEL 2 for the complex of ER and EST (1). MODEL 1 contains the entire LBD of the receptor protein,

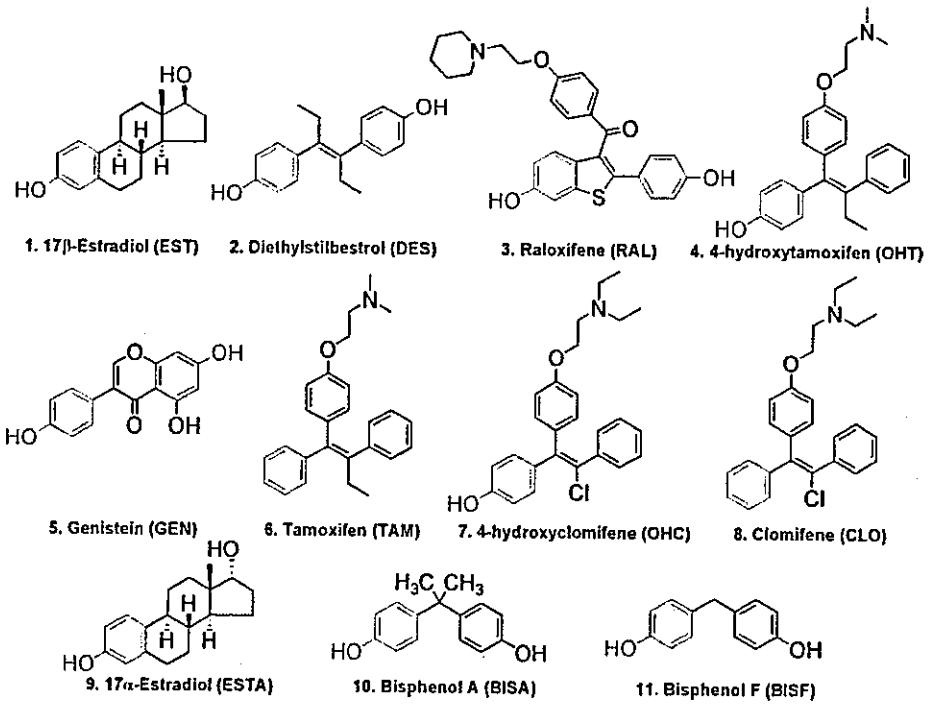

Figure 2. Estrogen-like compounds used in ER–ligand binding calculations.

consisting of 241 amino acid residues (residues 307–547). In addition to the challenge of calculating the bio-macromolecules, we are also interested in finding a reliable model for the efficient screening of ER–ligand docking, which plays an important role in the *in silico* process of drug discovery. Therefore, the large-scale calculations of MODEL 1 were performed only for four ligands, and most of the calculations were carried out using the efficient model, MODEL2. As suggested in Figure 1, the most important residues for ER–ligand binding are expected to be located in the vicinity of the ligand, where electrostatic and geometric interactions and hydrogen bondings seem to play the main role. We therefore anticipated the first-layered α-helices of the ligand to be a minimum model receptor, and MODEL 2 consists of a ligand, a water molecule, and 50 residues of ER around the ligand (residues 342–354, 382–395, 403–405, 417–429, and 520–526).

MODEL 3 was used for geometry optimizations of the hydrogen bond networks displayed in Figure 1. This model consists of a ligand, a water molecule, the side chains of the residues Glu353, Arg394, and His524, and the main chain of Leu387, and was further divided into two pieces, MODEL 3a and MODEL 3b, for the two ends of the ligand (Fig. 4). MODEL 3a consists of the ligand, Glu353, Arg394, Leu387, and a water molecule; MODEL 3b consists of the ligand and His524. The side chains of Glu353, Arg394, and His524 were truncated at Cγ, Cδ, and Cβ, respectively, and were capped with methyl groups. For MODEL 3a, the ligands were regarded as phenols with the exceptions for TAM (6) and CLO (8) that do not have hydroxy groups and were modeled by benzene. In MODEL 3b, the ligands were replaced by ROH : isopropyl alcohol for EST (1) and phenol for DES (2), RAL (3), and GEN (5) (the blue-colored parts of Fig. 2). All water mole-

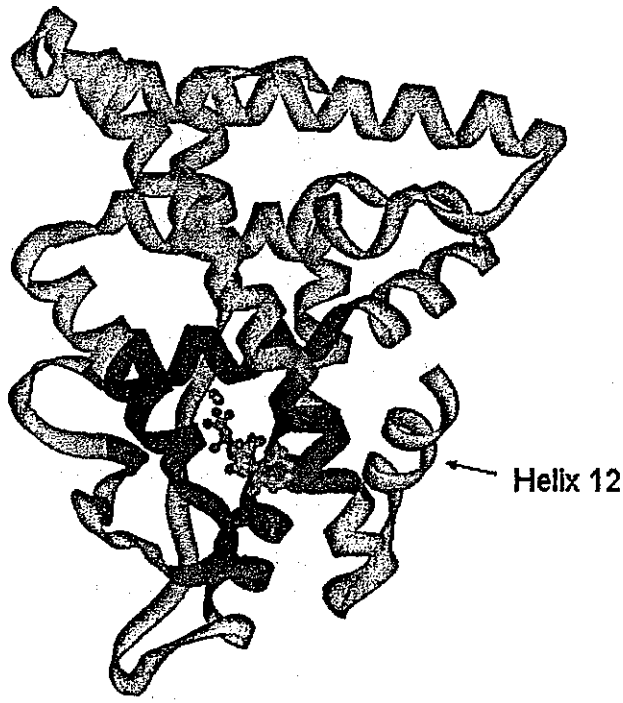

Figure 3. Ribbon display of the ERα LBD complexed with EST (1). MODEL 1, including 241 residues, is displayed as the whole complex, and in the inside, residues belonging to MODEL 2 (50 residues) are displayed as a purple ribbon surrounding the ligand (pink) and a water molecule (light blue). The position of Helix 12, which characterizes the agonism/antagonism of ER, is also indicated.

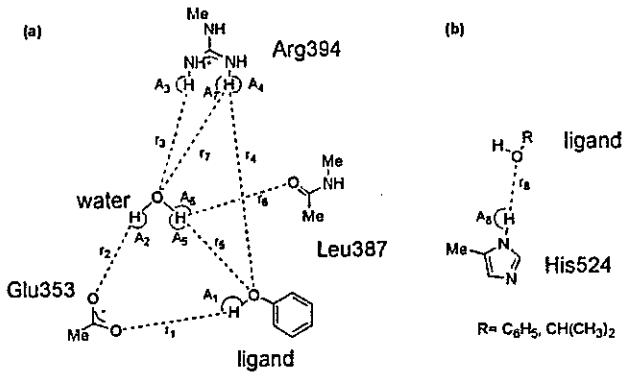

Figure 4. Structures of MODELs 3a and 3b used in geometry optimization of hydrogen bond networks. MODEL 3a, shown in (a), includes a portion of the ligand, a water molecule, Glu353, Arg394, and Leu387; MODEL 3b (b) includes a portion of the ligand and His524. Dotted lines indicate hydrogen bonds, and the distances (r_1–r_8) and the angles (A_1–A_8) are the hydrogen bond length (Å) and angles (degrees), respectively.

cules were eliminated except the one which directly mediates ER-ligand binding (Fig. 1).

Molecular Geometries

The geometries of the complexes were constructed based on the structural data obtained from the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB);[21] the entries are 1ERE, 3ERD, 1ERR, and 3ERT for the ERα-ligand complexes of EST (1), DES (2), RAL (3), and OHT (4), respectively.[12,13] Missing hydrogens, side chains, and main chains in the PDB files were complemented manually using the molecular graphic software InsightII.[41] Hydrogen atoms were added with both ends of the peptide chain and all the dissociative and associative residues in their charged states.

Eleven ER–ligand complexes were examined in this study. The initial atomic coordinates of the ER complexes with EST, RAL, DES, and OHT (1–4) were taken from the PDB files for 3ERE, 1ERR, 3ERD, and 3ERT, respectively. Because the coordinates for the ERα complexes with the other ligands (5–11) were not available in the PDB database, the initial binding geometries of these complexes were modeled as follows. First, geometries of ligands were modeled. The geometry of GEN (5) was taken from the PDB entry for 1QKM, which was the ERβ-GEN complex. TAM, OHC, and CLO (6–8) were modeled with the Insight II system[41] based on the geometry of OHT (4). TAM (6) was made by replacing the 4-hydroxy group of OHT (4) with H, and OHC (7) was made by replacing an ethyl group of OHT with the Cl atom, and the two methyl groups of OHT with two ethyl groups. CLO (8) was made by replacing the 4-hydroxy group of OHC (7) with H. The geometries of the other ligands, ESTA, BISA, and BISF (9–11), were optimized by conventional *ab initio* MO calculations at the HF/6-31G(d) level. Second, binding geometries of ER and ligand were modeled. For ESTA (9) and GEN (5), the positions were determined by superimposing the phenol groups, which colored red in Figure 2, on that of EST (1) in 1ERE. For the other ligands, either the phenol group or the phenyl group was superimposed on the phenol group of OHT (4) in 3ERT as well.

For the geometry optimization of the whole complexes, all the positions of hydrogen atoms, side chains, and backbones added in the previous procedure were optimized by CHARMM force field calculations[42] with the other heavy atoms fixed at the positions given in the PDB data. Then the geometries of the water molecule and hydrogen atoms of MODEL 3a and MODEL 3b that constitute the hydrogen bond network between ER and the ligand were optimized at the HF/6-31G(d) level. It appears that only the hydrogens of red and blue colored hydroxyl groups of ligands in Figure 2 and the hydrogen atoms of the water molecule were changed in this calculation.

Finally, to construct the structure of ER–ligand complexes with the optimized hydrogen bond network, the atomic coordinates of MODEL 1 and MODEL 2 were replaced with the corresponding optimized ones obtained from the MODEL 3 calculations.

Method of Calculations

Energy Calculations

The single-point energy calculations were carried out on MODEL 1 and MODEL 2 using the *ab initio* FMO method at the Hartree–Fock (HF) level with the STO-3G basis set (FMO-HF/STO-3G). Such large-scale calculations were achieved with the ABINIT-MP program (available from http://www.fsis.iis.u-tokyo.ac.jp/en/result/software), which was developed by our group for the calculations of bio-macromolecules.[35–39] The conventional HF method with the 6-31G(d) basis set was used for the energy calculations of MODEL 3 and the geometry optimization of some ligands, using the Gaussian98 program package.[43] Solvation energies of ligands were also calculated at HF/6-31G(d) with the Polarizable Continuum Model (PCM). The CHARMM force field calculations,[42] which is packaged in molecular graphic software InsightII,[41] were used for minimization of hydrogens and also for ER–ligand binding energies. In the binding energy calculations, ligands and their complexes were optimized with the harmonic atom constraint (force constant = 10.0) for backbone atoms, and then single-point energy calculations of them were performed without constrains.

The calculations were performed on a HITACHI SR8000 supercomputer at the Tsukuba Advanced Computing Center (TACC), on 16 Dual Pentium III 1-GHz clusters (32 CPUs), and on eight Dual Xeon 2.2-GHz clusters (16 CPUs). Most of the FMO calculations were done on the Dual Pentium III clusters. The elapsed time was ~6000 s for MOEL2 (50 residues) and ~50,000 s for MODEL 1 (241 residues).

The FMO Method

A brief description of the FMO method at the HF level is as follows. A molecule or a molecular cluster is divided into N fragments, and MO calculations on the fragments (monomers) and the fragment pairs (dimers) are performed to obtain the total energy and the properties of the system. In the following description, we assume that the monomers and the dimers are closed

shells. The Fock equation for the monomer I and the dimer IJ are solved,

$$\tilde{\mathbf{F}}^x\mathbf{C}^x = \mathbf{S}^x\mathbf{C}^x\tilde{\varepsilon}^x. \quad (1)$$

Here, $x = I$ for the monomer and $x = IJ$ for the dimer. $\tilde{\mathbf{F}}^x$ is a modified Fock matrix and, using the conventional notation, is written as

$$\tilde{\mathbf{F}}^x = \tilde{\mathbf{H}}^x + \mathbf{G}^x,$$
$$\tilde{H}^x_{\mu\nu} = H^x_{\mu\nu} + V^x_{\mu\nu} + \sum_i B_i\langle\mu|h_i\rangle\langle h_i|\nu\rangle, \quad (2)$$

where the one-electron Hamiltonian $\tilde{H}^x_{\mu\nu}$ is modified from the original one, $H^x_{\mu\nu}$ by adding the electrostatic potential V^x and the projection operator terms. The electrostatic potential consists of the nuclear attraction and two-electron term from the surrounding monomer K,

$$V^x_{\mu\nu} = \sum_{K(\neq x)} \left\{ \sum_{A\in K} \langle\mu|(-Z_A/|\mathbf{r} - \mathbf{r}_A|)|\nu\rangle + \sum_{\lambda\sigma\in K} D^K_{\lambda\sigma}(\mu\nu|\lambda\sigma) \right\}. \quad (3)$$

The projection operators are placed on atoms, where covalent bonds are detached for fragmentation, to divide and assign basis functions on the atoms into disconnected fragments.

The computational procedure of the FMO method is as follows. First, a set of the Fock equations [eq. (1)] for the monomers is solved repeatedly until all monomer densities become self-consistent. Second, the equations for the dimers are solved under the electrostatic potential from surrounding $(N-2)$ monomers. Finally, using the total energies of the monomer E_I and the dimer E_{IJ},

$$E_x = \frac{1}{2}\mathrm{Tr}\{\mathbf{D}^x(\tilde{\mathbf{H}}^x + \tilde{\mathbf{F}}^x)\} + E^{NR}_x, \quad (4)$$

the total energy of the system E is calculated by the following equation:

$$E = \sum_{I>J} E_{IJ} - (N-2)\sum_I E_I. \quad (5)$$

E^{NR}_x in eq. (4) is the nuclear repulsion energy.

In the practical calculations of the FMO method, one can save computational time by using the following approximations without losing significant accuracy. One is the Coulomb interaction approximation for well-separated dimers, which avoids to solve the HF equation (dimer-es). The others are related to the environmental electrostatic potentials defined in eq. (3): the fractional point charge approximations (esp-ptc) and the Mulliken orbital charge approximation (esp-aoc).[39] In this work, the dimer-es, the esp-ptc, and the esp-aoc approximations were applied to fragments whose separations were more than 2.0, 2.0, and 0.0, respectively, where the distances were given in units of the sum of the van der Waals radii of the closest contact atoms.

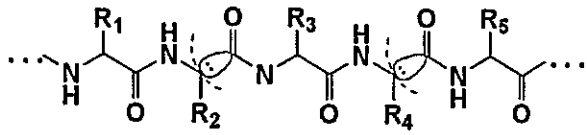

Figure 5. Fragmentation of peptides (broken arcs).

Fragmentation of Protein and Charge States of Amino Acid Residues

For FMO calculations, a molecule is divided into fragments. Generally, adoption of large fragments yields a total molecular energy approximating that from the conventional *ab initio* MO calculations but takes much computational time. The manner of dividing a molecule is thus a compromise between accuracy and computational time. It has been shown that the difference in total energies between the FMO and *ab initio* MO calculations at the HF/STO-3G level is less than 1 kcal/mol for several polypeptides and small proteins, when the fragmentation is done at α carbon atoms in blocks of two amino acid residues.[37] In this work, we applied this division scheme to the ER LBD protein (Fig. 5), expecting that the error in the total energy would be less than a few kcal/mol. The ligand molecules were not particularly large and thus were treated as a single fragment.

Several amino acid residues dissociate or associate with protons in aqueous solution. It is problematic to determine their charge states for electronic structural calculations. In this work, we assumed that the N-terminus and lysine and arginine residues were protonated and the C-terminus and aspartic and glutamic residues were deprotonated, although our calculations did not take the solvent into account. Thus, the total charge of the protein was −7e.

Results and Discussion

We calculated the binding energies between ER LBD and 11 ligands in Figure 2. Most estrogenic ligands have a phenol group, such as the steroid A ring, whose hydroxy group binds to the residues Glu353, Arg394, and Leu387 of ER through hydrogen bonds mediated by a single water molecule (Fig. 1). The endogenous steroidal ligand, EST (1), possesses the other hydroxy group on the D ring, which also makes a hydrogen bond with His524. The 11 ligands examined in this article can be classified into three types according to the number of hydrogen bonding sites. EST, DES, RAL, and GEN (1–3, 5) form hydrogen bonds with ER at two sites shown in red and blue in Figure 2, and are classified into type I. The intramolecular distances between the two hydroxy groups of the type I ligands are conserved around 11–12 Å, indicating that these ligands form consistent hydrogen bonds with ER. OHT (4) and OHC (7) have one hydrogen bonding phenol group and belong to type II. ESTA, BISA, and BISF (9–11) are also classified into type II despite their having two hydroxy groups, because these groups are structurally hindered from binding simultaneously. The hydrogen bonding sites of the type II ligands are also colored red in Figure 2. TAM (6) and CLO (8) do not have a hydroxy group and are classified into type III. The binding of these two ligands with ER might be weak because they cannot make a hydrogen bond.

Table 2. Optimized Bond Lengths (in Å) and Angles (in Degrees) Related to the Hydrogen Bond Network between the ER and Ligand.

	EST (1)	DES (2)	RAL (3)	OHT (4)	GEN (5)	TAM (6)	OHC (7)	CLO (8)	ESTA (9)	BISA (10)	BISF (11)
r_1	1.344	1.616	1.399	1.415	1.614	—	1.416	—	1.490	1.520	1.523
r_2	1.784	1.792	2.014	1.788	1.832	1.637	1.786	1.615	1.786	1.774	1.761
r_3	2.050	1.835	2.147	1.733	1.776	1.755	1.730	1.751	2.048	1.767	1.782
r_4	2.233	2.273	2.015	2.047	2.005	—	2.046	—	2.217	2.036	2.038
r_5	2.101	2.778	2.460	2.780	2.409	—	2.780	—	1.917	2.720	2.779
r_6	2.507	2.144	2.208	2.242	2.446	2.260	2.244	2.279	2.677	2.260	2.298
r_7	2.309	3.081	2.995	2.987	2.489	2.029	2.985	2.030	2.307	2.951	3.009
r_8	1.926	1.792	1.830	—	1.617	—	—	—	—	—	—
A_1	176.2	176.0	178.7	176.9	174.6	—	176.5	—	157.4	172.9	173.2
A_2	169.5	154.3	160.6	162.2	169.3	162.6	162.4	163.2	167.3	163.1	161.6
A_3	149.6	159.7	161.2	170.2	160.2	145.2	170.3	145.5	149.6	159.3	160.6
A_4	150.0	162.9	156.5	161.3	150.0	—	161.3	—	150.8	159.6	159.4
A_5	125.6	110.7	114.2	106.4	118.8	—	106.4	—	140.8	108.5	109.2
A_6	119.9	155.8	131.2	150.4	142.3	146.5	150.3	146.1	106.9	149.7	151.8
A_7	134.6	124.3	133.1	121.5	126.3	132.3	121.4	132.2	134.5	124.4	124.0
A_8	147.9	146.1	147.4	—	128.8	—	—	—	—	—	—

Structures of Hydrogen Bond Networks

Because the hydrogen bonds play a key role in the ER–ligand binding, it is very important to construct the proper model structures and to calculate the energy of hydrogen bonds with reasonable accuracy. We employed an *ab initio* quantum mechanical approach for the optimization of the hydrogen bond networks using a small model, MODEL 3. For the type I ligands, both MODELs 3a and 3b were used for describing the two binding sites. For the type II ligands, only MODEL 3a was used to optimize the hydrogen bond structure at one binding site. Concerning the type III ligands, we did not perform the optimizations of the structures, because these ligands do not make hydrogen bonds with the protein.

The optimized structure of the hydrogen bond network is shown by dotted lines in Figure 4. The geometrical parameters, the hydrogen bond lengths, r_1–r_8, and the angles, A_1–A_8, are given in Table 2. Glu353 and the phenol group of the ligand are strongly bound: the distances (r_1) are short, 1.3–1.6 Å, and the angles (A_1) are almost linear, 173–179°, except for ESTA (9). The water molecule is situated among Glu353, Arg394, Leu387, and the hydroxy group of the ligand (Fig. 4a). Normal hydrogen bonds are seen between the water molecule and Glu353 with 1.6–2.0 Å (r_2) and 155–170° (A_2), between the water molecule and Arg394 with 1.7–2.2 Å (r_3) and 145–170° (A_3), and between the ligand and Arg394 with 2.0–2.3 Å (r_4) and 150–160° (A_4). One of the hydrogen atoms of the water molecule is involved in both hydrogen bonds with the ligand (r_5: 1.9–2.8 Å and A_5: 105–140°) and with Leu387 (r_6 : 2.1–2.7 Å and A_6: 105–155°). Thus, the water molecule makes a bridge between the ligand and Leu387 through the hydrogen bonds. The hydrogen bond between Arg394 and the water molecule appears for the type III ligands, TAM (6) and CLO (8), with the bond length (r_7) of 2.03 Å and the bond angle (A_7) of 132°. Here, the hydrogen atom of Arg394 makes a hydrogen bond with the water molecule instead of with the hydroxy group of the type I and type II ligands. For MODEL 3b, hydrogen bonds are made between His524 and the other hydrogen bonding site of the type I ligands: the lengths are 1.6–1.9 Å (r_8) and the angles are 125–150° (A_8).

Binding Energies between ERα and Ligands

The binding energies were calculated at the FMO-HF/STO-3G levels, and for comparison, the CHARMM force field is also used. The energy of each of the three systems, that is, the receptor, E_{receptor}, the ligand, E_{ligand}, and the ER–ligand complex, E_{complex}, were calculated, where the hydrogen bonded water molecule was included in the receptor. The binding energy for a given ligand, ΔE_{ligand}, can be obtained as a difference in the energies of complex and of its components as follows,

$$\Delta E_{\text{ligand}} = E_{\text{complex}} - (E_{\text{receptor}} + E_{\text{ligand}}), \quad (6)$$

and the binding energies of a ligand relative to that of EST, $\Delta\Delta E_{\text{ligand}}$, are defined as

$$\Delta\Delta E_{\text{ligand}} = \Delta E_{\text{ligand}} - \Delta E_{\text{EST}}. \quad (7)$$

Then, the correlation between $\Delta\Delta E_{\text{ligand}}$ and the experimental RBA (Table 1) was examined. Here, we have ignored the geometry changes before and after binding, assuming that the binding energy changes resulted from geometrical relaxation of each ligand–ER complex are equivalent.

In the CHARMM results, the $\Delta\Delta E_{\text{ligand}}$ values were plotted against the experimental values of log(RBA/100) in Figure 6. For the eight compounds, 1–6, 9, and 10, no correlation was found between $\Delta\Delta E_{\text{ligand}}$ and log (RBA/100); the correlation coefficient r was 0.035. These calculations suggest that the intermolecular interactions between ER and ligands are poorly estimated by the CHARMM force field.

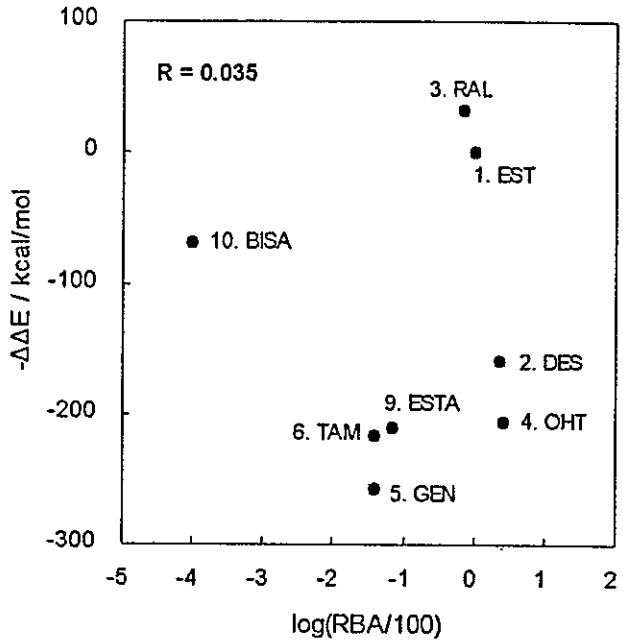

Figure 6. Relationship between calculated relative binding energies ($\Delta\Delta E$) vs. experimental RBA of eight ligands to ER LBD. Calculations were performed for MODEL 1 using the CHARMM force field.

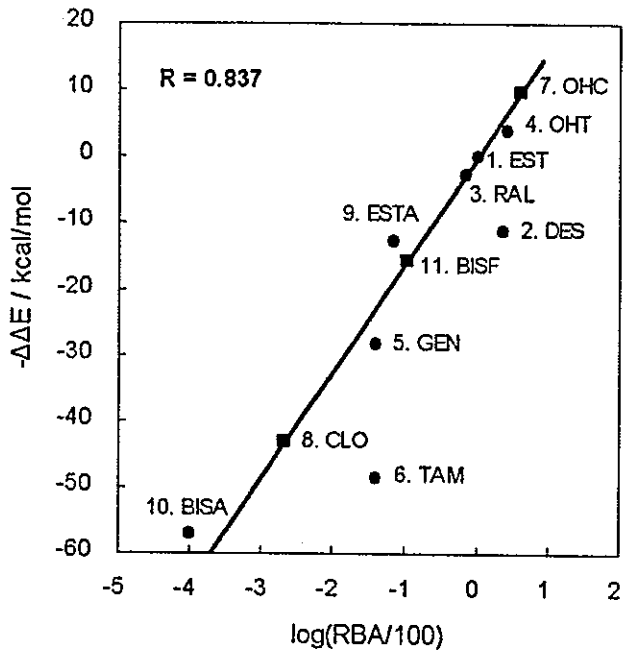

Figure 7. Relationship between calculated relative binding energies ($\Delta\Delta E$) vs. experimental RBA of 11 ligands to ER LBD. Calculations were performed using MODEL 2 at the FMO-HF/STO-3G level of theory.

The FMO results are described below. The $\Delta\Delta E_{ligand}$ values obtained from the MODEL 2 calculations are given in Table 3 and plotted against the experimental values in Figure 7. The ligands 1–6, 9, and 10, whose RBA values are known experimentally, are indicated by circles. Using the correlation equation obtained by the least-square fitting procedure, the log (RBA/100) values of ligands 7, 8, and 11, whose RBA are unknown, can be predicted from the calculated $\Delta\Delta E$ value; These are marked by squares. For the eight compounds, 1–6, 9, and 10, the correlation between $\Delta\Delta E$ and log (RBA/100) was good; the correlation coefficient r was 0.837. In particular, there was a very good correlation ($r = 0.931$) for seven of the eight ligands, with TAM (6) being the exception.

TAM (6) is an outlier of the correlation line. This may be due to the artificially imposed orientation of the ligand in the active site resulting from a superimposition of TAM on OHT (4) in the crystal structure. TAM cannot form a hydrogen bond with ER due to lack of a hydroxy group. There is thus no reason to restrict the ligand in the active site in the same manner as other hydrogen bonding ligands such as OHT. Despite that, we were simply substituting TAM for OHT in the crystal structure. This might bring artificial instability of the calculated binding energy and shift the position in the plot lower than where it should be in Figure 7. For the same reason, the predicted RBA of CLO (8) should be shifted upward to some extent. Regarding BISA (10), although the plot is almost in the correlation line, the repulsive binding energy implies low reliability of the calculated energy. To avoid these deficiencies, it is necessary to optimize the structure of the ligand and the surrounding residues together. This is, however, almost impossible at the present time, because too much computer time is required.

Using MODEL 1, the $\Delta\Delta E$ values are also calculated for the complexes of ER with the ligands EST, DES, RAL, and OHT (1–4). The calculated binding energies (ΔE) are also shown in Table 3. The ΔE values of MODEL 1 are almost identical with those of MODEL 2. The differences in ΔE between the two models are below 3 kcal/mol for the complexes of EST (1), DES (2), and

Table 3. Experimental RBA, Calculated Binding Energies (ΔE), and Relative Binding Energies ($\Delta\Delta E$) of Several Ligands with ER.

		MODEL2 (50 residues)		MODEL1 (241 residues)	
Ligand	Log(RBA/100)	ΔE	$\Delta\Delta E$	ΔE	$\Delta\Delta E$
EST (1)	0.00	−37.80	0.00	−37.65	0.00
DES (2)	0.37	−26.70	11.10	−28.33	9.32
RAL (3)	−0.16	−35.30	2.50	−26.13	11.52
OHT (4)	0.41	−41.73	−3.92	−38.19	−0.54
GEN (5)	−1.40	−9.42	28.38	—	—
TAM (6)	−1.40	10.62	48.42	—	—
OHC (7)	—	−47.62	−9.82	—	—
CLO (8)	—	5.29	43.10	—	—
ESTA (9)	−1.15	−25.07	12.73	—	—
BISA (10)	−4.00	19.22	57.02	—	—
BISF (11)	—	−22.23	15.57	—	—

Energies are in kcal/mol.

Table 4. Solvation Free Energy of Ligands (G^{sol}_{ligand}).

Ligand	G^{sol}_{ligand}	ΔG^{sol}_{ligand}
EST (1)	−13.07	0.00
DES (2)	−12.17	0.90
RAL (3)	−19.26	−6.19
OHT (4)	−10.75	2.32
GEN (5)	−18.69	−5.62
TAM (6)	−4.69	8.38
OHC (7)	−10.56	2.51
CLO (8)	−4.99	8.08
ESTA (9)	−11.07	2.00
BISA (10)	−10.96	2.11
BISF (11)	−12.24	0.83

They are energy differences between ligands in gas phase (E^{gas}_{ligand}) and in water of PCM (G^{PCM}_{ligand}), and the relative solvation free energies of ligand (ΔG^{sol}_{ligand}) were calculated relative to that of EST (ΔG^{sol}_{EST}).

OHT (4). In the case of RAL (3), however, a rather large discrepancy is observed between the two models. This is thought to be due to the fact that Leu536 and Leu539, which are not included in MODEL 2, are located close enough to interact with the ligand, which is not the case for the other three ligands. Thus, MODEL 2 with the 50 residues is a reliable model for describing the interaction between ER and the ligands, but for a large ligand such as RAL, additional residues which make contact with the ligand should be included.

There would be considerable contribution of many effects to the binding affinity, which were ignored in our calculations. A typical one is the solvent effect. To consider the solvent effect on ER–ligand binding, we estimated solvation free energies of ligands (G^{sol}_{ligand}) using the PCM model as a difference between ligand energies in the gas phase (E^{gas}_{ligand}) and ligand free energies in water of the PCM model (G^{PCM}_{ligand});

$$G^{sol}_{ligand} = G^{PCM}_{ligand} - E^{gas}_{ligand}, \quad (8)$$

$$\Delta G^{sol}_{ligand} = G^{sol}_{ligand} - G^{sol}_{EST}. \quad (9)$$

As shown in Table 4, the differences of solvation free energies between EST(1) and other ligands were calculated to be less than 8.4 kcal/mol, and the ligand sensitivity on the solvent effect of the relative binding energy in eq. (7) is expected to have a less order of magnitude because the cancellation of solvation energies would occur among $G^{sol}_{complex}$, $G^{sol}_{receptor}$ and G^{sol}_{ligand} in binding energy calculations synonymous with eq. (6). Another effect, that is, hydrophobic interaction and induced fitting, were ignored and the dispersion energy is not included in the HF calculations. Our calculated results nevertheless obtained a nice correlation with the experimental results, and therefore, the relative binding affinity is thought to be correlated with the enthalpic relative binding energies, and other effects for each ligand could be assumed to be similar.

Charge Distribution

The difference in the net charges between a complex and individual component molecules is shown for several residues, the water molecule, and the ligands in Table 5 and Table 6 for MODEL 2 and MODEL 1, respectively. Because these two models give similar results, the following discussions are given based on MODEL 2. In ER–ligand complexes, the total net charges of ligands are negative by −0.001~−0.181e, and of the same order of positive charges are induced on Glu353 by +0.006~+0.198e. The changes in the total net charges of Leu387, Arg394, His524, and the water molecule are very small. The type III ligands, which make no hydrogen bonds with ER, show very little change in their charge states. Therefore, considerable electrons are supplied from GLU353 to the ligands with the exception of the nonhydrogen binding ones, which is consistent with the fact that a strong hydrogen bond is formed between these ligands (proton donors) and GLU353 (a proton acceptor).

The number of electrons transferred from ER to the ligands is highly related with the binding energy; ΔE becomes larger or the complex becomes more stable with the increase in the negative charge of the ligand (Table 5). Thus, most of the stabilization in the ER-ligand docking arises from the ligand–Glu353 interaction. This fact suggests that the hydrogen bond between ER and the ligand plays an important role in characterizing the charge transfer and the concerted stabilization of the ER–ligand complex. This does not mean that a simple model that assumes the additivity of individual hydrogen bonds between the ligand and the residues suffices for the description of the interactions between the protein and the ligand. We note that not only the hydrogen bonds but also the electrostatic interactions between the ligand and the surrounding residues contribute to the protein–ligand docking.

Conclusions

We have calculated the binding energies between ER and its 11 ligands using the *ab initio* FMO method, which allows the calculation of large molecules. We took particular note of the hydrogen bond network formed between ER and the ligands. Three models of the protein were used for the description of the proper hydrogen bonding network: MODEL 1, MODEL 2, and MODEL 3, respectively consisting of 241, 50, and 4 amino acid residues. Although the classical CHARMM force field calculations for the entire ER LBD gave poor correlation, the relative binding energies obtained from the MODEL 2 FMO calculations were in good correlation with the experimental RBA (logRBA), with a correlation coefficient of 0.837. These results show the advantage of FMO calculations, and suggest that the ER–ligand interaction is localized in the binding region and is properly described by considering the amino acid residues in the first layered α-helices of the ligand. The entire ER LBD should be treated for the study of postbinding, including repositioning of Helix12 due to the binding of agonists or antagonists. It was also found that the binding energy was related to charges transferred from the protein to the ligand upon the complexization; as the transferred charge increased, the binding energy also became larger.

The methods presented herein may provide a powerful tool for assessing the affinity of putative xenoestrogens *in silico* prior to biological experiments. Our results show that the FMO method and *ab initio* quantum mechanical calculations are efficient and valid tools for predicting the binding affinities of ligands to proteins. However, further development of quantum mechanical

Table 5. Charge Differences Between ER–Ligand Complexes ("$LR_{complex}$," MODEL2) and Individual Molecules ("L + R").

	GLU353			LEU387			ARG394			HIS524			Water			Ligand			
	$LR_{complex}$	L + R	Δq	$LR_{complex}$	L + R	Δq	$LR_{complex}$	L + R	Δq	$LR_{complex}$	L + R	Δq	$LR_{complex}$	L + R	Δq	$LR_{complex}$	L + R	Δq	ΔE
OHC (7)	−0.779	−0.976	0.198	0.007	0.009	−0.002	0.772	0.793	−0.022	−0.013	−0.014	0.001	−0.017	−0.015	−0.003	−0.179	0.000	−0.179	−47.62
OHT (4)	−0.781	−0.976	0.195	0.008	0.009	−0.001	0.772	0.793	−0.022	−0.014	−0.014	0.000	−0.017	−0.015	−0.003	−0.181	0.000	−0.181	−41.73
EST (1)	−0.778	−0.973	0.195	0.002	0.002	0.001	0.783	0.793	−0.010	−0.050	−0.032	−0.018	−0.054	−0.047	−0.007	−0.169	0.000	−0.169	−37.80
RAL (3)	−0.821	−1.019	0.198	0.007	0.009	−0.002	0.709	0.731	−0.022	−0.062	−0.039	−0.024	−0.029	−0.028	−0.002	−0.131	0.000	−0.131	−35.30
DES (2)	−0.859	−0.982	0.123	0.022	0.021	0.002	0.786	0.796	−0.010	−0.064	−0.035	−0.028	−0.015	−0.011	−0.004	−0.089	0.000	−0.089	−26.70
ESTA (9)	−0.839	−0.974	0.136	0.005	0.003	0.002	0.783	0.793	−0.010	−0.031	−0.031	0.000	−0.061	−0.044	−0.017	−0.122	0.000	−0.122	−25.07
BISF (11)	−0.880	−0.977	0.097	0.011	0.009	0.002	0.771	0.794	−0.022	−0.015	−0.014	−0.001	−0.018	−0.015	−0.003	−0.075	0.000	−0.075	−22.23
GEN (5)	−0.857	−0.981	0.124	−0.004	−0.002	−0.002	0.748	0.766	−0.018	−0.060	−0.031	−0.029	−0.013	−0.012	−0.001	−0.076	0.000	−0.076	−9.42
CLO (8)	−0.970	−0.976	0.006	0.005	0.009	−0.004	0.791	0.793	−0.002	−0.013	−0.014	0.001	−0.015	−0.015	0.000	−0.001	0.000	−0.001	5.29
TAM (6)	−0.970	−0.976	0.006	0.006	0.009	−0.003	0.791	0.793	−0.002	−0.013	−0.014	0.000	−0.015	−0.015	0.000	−0.006	0.000	−0.006	10.62
BISA (10)	−0.880	−0.977	0.097	0.011	0.009	0.002	0.771	0.793	−0.022	−0.015	−0.014	0.000	−0.018	−0.015	−0.003	−0.063	0.000	−0.063	19.22

$\Delta q = q(\text{complex}) - q(\text{individual})$. Units are in a.u.

Table 6. Charge Differences Between ER–Ligand Complexes ("$LR_{complex}$," MODEL 1) and Individual Molecules ("L + R").

	GLU353			LEU387			ARG394			HIS524			Water			Ligand			
	$LR_{complex}$	L + R	Δq	$LR_{complex}$	L + R	Δq	$LR_{complex}$	L + R	Δq	$LR_{complex}$	L + R	Δq	$LR_{complex}$	L + R	Δq	$LR_{complex}$	L + R	Δq	ΔE
OHT (4)	−0.737	−0.919	0.182	−0.013	−0.014	0.000	0.861	0.884	−0.023	0.017	0.017	0.000	−0.010	−0.012	0.002	−0.175	0.000	−0.175	−38.19
EST (1)	−0.741	−0.923	0.182	−0.012	−0.014	0.002	0.903	0.912	−0.010	−0.047	−0.022	−0.025	−0.053	−0.046	−0.008	−0.151	0.000	−0.151	−37.65
RAL (3)	−0.809	−0.998	0.189	−0.010	−0.009	−0.002	0.843	0.867	−0.024	−0.052	−0.027	−0.024	−0.028	−0.026	−0.002	−0.121	0.000	−0.121	−26.13
DES (2)	−0.826	−0.938	0.112	0.008	0.005	0.003	0.900	0.910	−0.010	−0.069	−0.038	−0.031	−0.014	−0.010	−0.004	−0.078	0.000	−0.078	−28.33

$\Delta q = q(\text{complex}) - q(\text{individual})$. Units are in a.u.

methods will be needed to obtain more reliable binding energies from the calculations. It will be necessary to allow optimization of the geometry of entire complexes, particularly induced-fit complexes, and to enter the effects of the solvent into the calculations. We are currently developing the FMO method along these lines.

Acknowledgments

K.F. thanks Emeritus Professor Toshio Fujita of Kyoto University for his valuable input. This work was supported by a grant (MF-16) from the Organization for Pharmaceutical Safety and Research. A part of this research was done in conjunction with the "Frontier Simulation Software for Industrial Science (FSIS)" project supported by the IT program of the Ministry of Education, Culture, Sports, Science, and Technology (MEXT). Part of the calculations were carried out on a HITACHI SR8000 supercomputer at the Tsukuba Advanced Computing Center (TACC) of the National Institute of Advanced Industrial Science and Technology (AIST).

References

1. Kuiper, G. G. J. M.; Enmark, E.; Pelto-Huikko, M.; Nilsson, S.; Gustafsson, J.-Å. Proc Natl Acad Sci USA 1996, 93, 5925.
2. Peach, K.; Webb, P.; Kuiper, G. G. J. M.; Nilsson, S.; Gustafsson, J.-Å.; Kushner, P. J.; Scanlan, T. S. Science 1997, 277, 1508.
3. Kuiper, G. G. J. M.; Carlsson, B.; Grandien, K.; Enmark, E.; Häggblad, J.; Nilsson, S.; Gustafsson, J.-Å. Endocrinology 1997, 138, 863.
4. Kuiper, G. G. J. M.; Lemmen, J. G.; Carlsson, B.; Corton, J. C.; Safe, S. H.; van der Saag, P. T.; van der Burg, B.; Gustafsson, J.-Å. Endocrinology 1998, 139, 4252.
5. Barkhem, T.; Carlsson, B.; Nilsson, Y.; Enmark, E.; Gustafsson, J.-Å.; Nilsson, S. Mol Pharmacol 1998, 54, 105.
6. Mangelsdorf, D. J.; Thummel, C.; Beato, M.; Herrlich, P.; Schütz, G.; Umesono, K.; Blumerg, B.; Kastner, P.; Mark, M.; Chambon, P.; Evans, R. M. Cell 1995, 83, 835.
7. Nuclear Receptors Nomenclature Committee. Cell 1999, 97, 161.
8. Glass, C. K.; Rosenfeld, M. G. Genes Dev 2000, 14, 121.
9. Renaud, J. P.; Moras, D. Cell Mol Life Sci 2000, 57, 1748.
10. Steinmetz, A. C. U.; Renaud, J.-P.; Moras, D. Annu Rev Biophys Biomol Struct 2001, 30, 329.
11. Kato, S. Mol Med 2000, 37, 1170.
12. Brzozowski, A. M.; Pike, A. C.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engström, O.; &Oulm;hman, L.; Greene, G. L.; Gustafsson, J.-Å.; Carlquist, M. Nature 1997, 389, 753.
13. Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. Cell 1998, 95, 927.
14. Ekena, K.; Weis, K. E.; Katzenellenbogen, J. A.; Katzenellenbogen, B. S. J Biol Chem 1997, 272, 5069.
15. Nilsson, S.; Kuiper, G.; Gustafsson, J.-Å. Trends Endocrinol Metab 1998, 9, 387.
16. Sonnenschein, C.; Soto, A. M. J Steroid Biochem Mol Biol 1998, 65, 143.
17. Fang, H.; Tong, W.; Perkins, R.; Soto, A. M.; Prechtl, N. V.; Sheehan, D. M. Environ Health Perspect 2000, 108, 723.
18. Tanenbaum, D. M.; Wang, Y.; Williams, S. P.; Sigler, P. B. Proc Natl Acad Sci USA 1998, 95, 5998.
19. Pike, A. C. W.; Brzozowski, A. M.; Hubbard, R. E.; Bonn, T.; Thorsell, A.-G.; Engström, O.; Ljunggren, J.; Gustaffson, J.-Å; Carlquist, M. EMBO J 1999, 18, 4608.
20. Gangloff, M.; Ruff, M.; Eiler, S.; Duclaud, S.; Wurtz, J. M.; Moras, D. J Biol Chem 2001, 276, 15059.
21. The RCSB Protein Data Bank (http://www.rcsb.org/.). Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235.
22. Gantchev, T. G.; Ali, H.; van Lier, J. E. J Med Chem 1994, 37, 4164.
23. Bradbury, S. P.; Mekenyan, O. G.; Ankley, G. T. Environ Toxicol Chem 1996, 15, 1945.
24. Waller, C. L.; Oprea, T. I.; Chae, K.; Park, H.-K.; Korach, K. S.; Laws, S. C.; Wiese, T. E.; Kelce, W. R.; Gray, L. E., Jr. Chem Res Toxicol 1996, 9, 1240.
25. Tong, W.; Perkins, P.; Xing, L.; Welsh, W. J.; Sheehan, D. M. Endocrinology 1997, 138, 4022.
26. Bradbury, S. P.; Mekenyan, O. G.; Ankley, G. T. Toxicol Chem 1998, 17, 15.
27. Oostebrink, B. C.; Pitera, J. W.; van Lipzig, M. M. H.; Meerman, J. H. N.; van Gunsteren, W. F. J Med Chem 2000, 43, 4594.
28. Kirchhoff, P. D.; Brown, R.; Kahn, S.; Waldman, M.; Venkatachalam, C. M. J Comput Chem 2001, 22, 993.
29. Sato, F.; Yoshihiro, T.; Era, M.; Kashiwagi, H. Chem Phys Lett 2001, 341, 645.
30. Challacombe, M.; Schwegler, E. J. J Chem Phys 1997, 106, 5526.
31. Van-Alsenoy, C.; Yu, C.-H.; Peerers, A.; Martin, J. M. L.; Schäfer, L. J Phys Chem 1998, A102, 2246.
32. Sato, F.; Yoshihiro, T.; Okazaki, I.; Kashiwagi, H. Chem Phys Lett 1999, 310, 523.
33. Scuseria, G. E. J Phys Chem 1999, A103, 4782.
34. Tsuda, K.; Kaneko, H.; Shimada, J.; Takada, T. Comp Phys Commun 2001, 142, 140.
35. Kitaura, K.; Sawai, T.; Asada, T.; Nakano, T.; Uebayasi, M. Chem Phys Lett 1999, 312, 319.
36. Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. Chem Phys Lett 1999, 313, 701.
37. Nakano, T.; Kaminuma, T.; Sato, T.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. Chem Phys Lett 2000, 318, 614.
38. Kitaura, K.; Sugiki, S.; Nakano, T.; Komeiji, Y.; Uebayasi, M. Chem Phys Lett 2001, 336, 163.
39. Nakano, T.; Kaminuma, T.; Sato, T.; Fukuzawa, K.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. Chem Phys Lett 2002, 351, 475.
40. Fukuzawa, K.; Kitaura, K.; Nakata, K.; Kaminuma. K.; Nakano, T. Pure Appl Chem 2003, 75, 2405.
41. InsightII Version 98.0, Molecular Simulations Inc., San Diego, CA, 1998.
42. CHARMm: Chemistry at HARvard Macromolecular Mechanics (CHARMm), Version 25.2 Revision: 98.0731.; Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J Comp Chem 1983, 4, 187; MacKerell, A. D., Jr.; Brooks, B.; Brooks, C. L., III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. The Encyclopedia of Computational Chemistry, Schleyer, P. v. R., et al., Eds.; John Wiley & Sons: Chichester, 1998; p. 271, vol. 1.
43. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. Gaussian 98, Revision A.7; Gaussian, Inc.: Pittsburgh, PA, 1998.

Pharmacogenomics in Drug Transporters: Functional Analysis of Genetic Polymorphisms

薬物トランスポーターのファーマコゲノミクス：遺伝的多型の機能解析

Toshihisa Ishikawa, PhD Professor, Department of Biomolecular Engineering
Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology*

石川 智久 東京工業大学 大学院生命理工学研究科
生体分子機能工学専攻 教授

Dr. T. Ishikawa

Keywords:
Single nucleotide polymorphism (SNP)
Drug transporter
Pharmacogenomics
Pharmacokinetics

1塩基多型(SNP)
薬物トランスポーター
ファーマコゲノミクス
薬物動態

Summary

Evidence is accumulating to strongly suggest that drug transporters are one of the determinant factors governing the pharmacokinetic profile of drugs. Hitherto a variety of drug transporter genes have been cloned and classified into either solute carriers (SLC) or ATP-binding cassette (ABC) transporters. Such drug transporters are expressed in various tissues such as the intestine, brain, liver, and kidney, to play critical roles in the absorption, distribution and excretion of drugs. However, at the present time, information is limited regarding the genetic polymorphism of drug transporters and its impact on the function. In this context, we have undertaken the functional analyses of the polymorophisms identified in drug transporter genes. This review addresses part of our most recent studies to exemplify the importance of genetic polymorphisms in drug transporters.

近年，薬物トランスポーターが薬物の体内動態を規定する重要な因子であるという研究証拠が蓄積しつつある。これまで数多くの薬物トランスポーター遺伝子がクローニングされて，それらは溶質キャリアー(SLC, solute carrier)またはABC(ATP-binding cassette)トランスポーターに区分されている。これら薬物トランスポーターは小腸，脳，肝臓，腎臓などさまざまな臓器／器官に発現して，薬物の吸収・分布・排泄において重要な役割を担っている。しかしながら現在のところ，薬物トランスポーターの遺伝子多型および機能へのインパクトに関する情報はまだそれほど多くはない。我々は，薬物トランスポーターの遺伝子多型の機能を解明し，薬の効果・副作用に関連する知見を得ることに着手した。この総説では，その研究成果の一部を紹介する。

* Nagatsuta 4259, Midori-ku,Yokohama-shi, Kanagawa, 226-8501, Japan Tel: +81-(0)45-924-5800 Fax: +81-(0)45-924-5838
E-mail: tishikaw@bio.titech.ac.jp Website: http://www.ishikawa-lab.bio.titech.ac.jp/

Introduction

In the last decade of the 20th century, the development of high throughput screening and combinatorial chemistry technologies accelerated the drug discovery process. In the 21st century, emerging genomic technologies (i.e., bioinformatics, functional genomics, and pharmacogenomics) are shifting the paradigm of drug discovery research and improving the strategy of medical care for patients. Identifying human DNA sequences, genomic structures, and human genetic variations, along with changes in gene and protein expression allows researchers and clinicians to more precisely define diseases and, in turn, to achieve the goal of "personalized medicine".

In order to realize the personalized medicine, it is critically important to understand molecular mechanisms underlying inter-individual differences in the drug response, namely, pharmacological effect vs. side effect [1]. The occurrence of the variations among persons in the drug response may involve many different causes, for example, genetic variations and/or expression levels of drug target molecules including membrane receptors, nuclear receptors, signal transduction components, enzymes, etc. as well as those of drug metabolizing enzymes and drug transporters (**Fig. 1**). Observations of inter-individual variations in different drug responses have led to the development of pharmacogenetics and pharamacogenomics.

Genetic polymorphisms in drug response-related genes

Drug transporters and drug-metabolizing enzymes are important because they play pivotal roles in determining the pharmacokinetic profiles of drugs and, by extension, their overall pharmacological effects (i.e., drug absorption, drug distribution, drug metabolism and elimination, drug concentration at the target site, and the number and morphology of target receptors). The effects of drug transporters on the pharmacokinetic profile of a drug depend on their expression and functionality. Indeed, the expression of drug transporters can be modulated by endogenous and exogenous factors, including drugs, themselves. It is also now known that inherited differences among individuals may also affect drug efficacy and toxicity. Such inherited differences include genetic polymorphisms in drug targets and drug-metabolizing enzymes, as well as in drug transporters. Hitherto, pharmacogenetics, the field dealing with such inherited differences and their effect on pharmacokinetics, has significantly contributed to our understanding of genetic causes underlying differences in drug metabolism (e.g., cytochrome P-450 mediated drug metabolism). In fact, recent technological advances allowing massive molecular sequencing have in turn allowed us to identify single nucleotide polymorphisms (SNPs) as one possible cause of variable drug response among individuals [2,3]. In light of such advances, it is important to carefully examine the clinical significance, if any, of polymorphisms in drug response genes, including drug transporters.

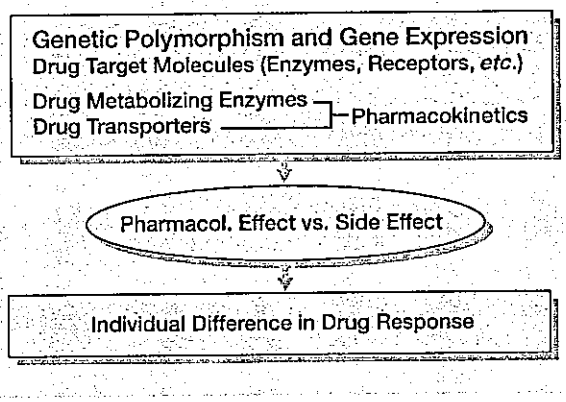

Fig. 1: Impact of the genetic polymorphism and/or gene expression of drug targets, drug metabolizing enzymes, and drug transporters on the drug response.

SLC and ABC transporter families

There is accumulating evidence to strongly suggest that drug transporters are one of the determinant factors governing the pharmacokinetic profile of drugs. Indeed, drug transporters are expressed in various tissues such as the intestine, brain, liver, and kidney, to play critical roles in the absorption, distribution and excretion of drugs. Hitherto, a variety of drug transporters have been cloned, and remarkable progress has been made in charactering the molecular properties and functions of these transporters. Such transporters have been classified as either primary or secondary active transporters. The primary active transporters include ATP-binding cassette (ABC) transporters that utilize the ATP hydrolysis as the driving force for solute transport [4,5]. On the other hand, the secondary transporters, e.g., many of solute carrier (SLC) transporters, are driven by an exchange of intra/extra-cellular ions [6,7]. Each gene family of transporters comprises of a multiplicity of members. The human Gene Nomenclature Committee has classified transporters by standardized names such as the SLC family and the ABC transporters. **Table 1** summarizes major drug transporters expressed in small intestine, blood brain barrier, liver and kidney.

The functions and substrate specificities of drug transporters have been characterized by several *in vitro* and *in vivo* techniques using cells expressing the transporter gene or using gene-knockout animals. In particular, construction of *in vitro* expression systems using human transporter cDNA clones provides as useful models to evaluate the substrate specificity. In addition, tissue distribution and levels of expression of the drug transporters convey important information for the prediction of the *in vivo* pharmacokinetic profile of drugs.

There are many factors that can affect the function as well as the expression of drug transporters. Those factors may involve genetic mutations, SNPs, splicing, transcriptional regulation, stability of mRNA, post-translational modification, and intracellular localization (**Fig. 2**). Evaluation of such factors is critically important to understand the whole picture of pharamacogenomics of drug transporters. Functional analysis of the polymorphism of drug transporters is one of such important approaches.

Organ/Tissue	ABC Transporter	SLC Transporter		
		Peptide transporter	Anion transporter	Cation transporter
Small intestine	ABCB1 (P-gp/MDR1) ABCB4 (MDR2) ABCC2 (cMOAT/MRP2) ABCC3 (MRP3) ABCC4 (MRP5) ABCC5 (MRP5) ABCC6 (MRP6) ABCG2 (BCRP/MXR/ABCP)	PEPT1 (SLC15A1)	MCT1 (SLC16A1) MCT4 (SLC16A4) MCT5 (SLC16A5) MCT8 (SLC16A8) OATP-B (SLC21A9) OATP-D (SLC21A11) OATP-E (SLC21A12) PGT (SLC21A2) AE2 (SLC4A2) ASBT (SLC10A2)	OCT1 (SLC22A1) OCTN1 (SLC22A4) OCTN2 (SLC22A5)
Blood brain barrier	ABCB1 (P-gp/MDR1) ABCC1 (MRP1) ABCG2 (BCRP/MXR/ABCP)		MTC1 (SLC16A1) MCT2 (SLC16A2) OAT1 (SLC22A6) OAT3 (SLC22A8) OATP-A/OATP (SLC21A3)	OCTN2 (SLC22A5) OCT2 (SLC22A1) OCT3 (SLC22A5)
Liver	ABCB1 (P-gp/MDR1) ABCB4 (MDR2) ABCB11 (SPGP/BSEP) ABCC2 (cMOAT/MRP2) ABCC3 (MRP3) ABCG2 (BCRP/MXR/ABCP)		OATP-B (SLC21A9) OATP-C/LST-1 (SLC21A6) OATP-8 (SLC21A8) NPT1 (SLC17A1)	OCT1 (SLC22A1) OCTN2 (SLC22A5) NTCP (SLC10A1)
Kidney	ABCB1 (P-gp/MDR1) ABCC1 (MRP1) ABCC2 (MRP2)	PEPT1 (SLC15A1) PEPT2 (SLC15A2)	OAT1 ((SLC22A6) OAT3 (SLC22A8) OAT4 (SLC222A9) NPT1 (SLC17A1)	OCT1 (SLC22A1) OCT2 (SLC22A2) OCT3 (SLC22A3) OCTN1 (SLC22A4) OCTN2 (SLC22A5)

Table 1: ABC and SLC Transporters expressed in small intestine, blood brain barrier, liver and kidney

On 6 September 2000, the 43 members of the Japan Pharmaceutical Manufacturers Association (JPMA) established the Pharma SNP consortium (PSC) to conduct research into pharmacokinetic gene polymorphism in the Japanese population[8]. During the period of 2000 to 2003, using blood samples donated by about 1000 Japanese volunteers, the PSC identified SNP in approximately 180 pharmacokinetic genes including drug metabolizing enzymes and drug transporters. The PSC project then created a database of SNPs and information from expression and functional analyses of protein variants. The overall objective is to gather common fundamental data necessary for research into drug responsiveness in the Japanese population. In the PSC project, we have undertaken the functional analyses of SNPs discovered in drug transporter genes[9]. **Fig. 3** summarizes the strategy of our functional analysis. Accordingly, this review addresses part of our recent studies to exemplify the importance of genetic polymorphisms in drug transporters.

Naturally occurring SNPs in *ABCB1* gene

ABCB1 (P-glycoprotein/MDR1) is gaining attention for its involvement in drug absorption by the small intestine and drug penetration into the brain; it is expressed in a variety of normal cells and organs, and its modulation in these tissues can influence the activity and bioavailability of drugs. In the intestine, for instance, modulation of ABCB1 may control the degree of drug uptake after drug ingestion. At the blood-brain barrier, high P-glycoprotein levels can limit the uptake of desired drugs into the brain; conversely, low ABCB1 activity can lead to abnormally increased accumulation and undesirable side effects.

To date, genetic variations of the human *ABCB1* gene have been most extensively studied. Hitherto about 50 SNPs and 3 insertion/deletion polymorphisms in the *ABCB1* gene have been reported[10-13]. In addition, twelve novel SNPs of ABCB1 were reported in Japanese patients with ventricular tachycardia who were administered amiodarone[14]. Several preclinical and clinical studies have provided evidence for the naturally occurring polymorphisms in ABCB1 and their effects on drug absorption, distribution and elimination. Hoffmeyer *et al.*[15] first reported multiple polymorphisms in the

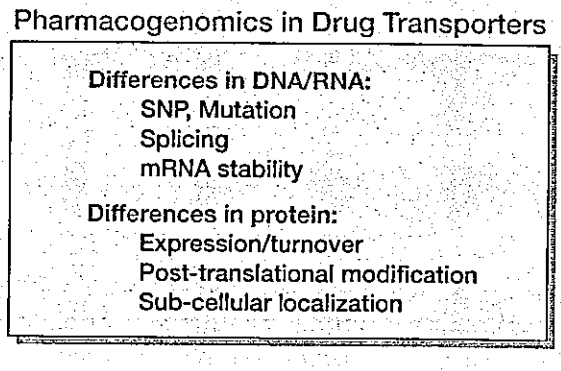

Fig. 2: Potential factors that affect the pharmacogenomics of drug transporters.

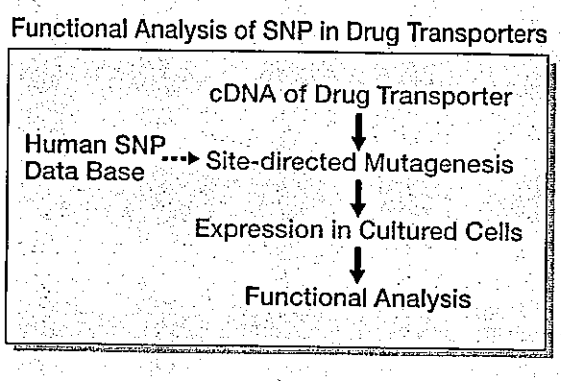

Fig. 3: Strategy for the functional analysis of non-synonymous SNPs in drug transporters.