Chem-Bio Informatics Journal, Vol. 4, No. 4, pp.121-132(2004)

A Statistical Analysis of an Effective Method to Conduct *In Silico* Screening for Active Compounds

Keiji Kakumoto[1, 2] *, Shota Yamanaka[1], Chikuma Hamada[2], and Isao Yoshimura[2]

[1]*Otsuka Pharmaceutical Co., Ltd.,*
463-10 Kagasuno, Kawauchi-cho, Tokushima 771-0192, Japan
[2] *Graduate School of Engineering, Tokyo University of Science,*
1-3 Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan

**E-mail: k_kakumoto@research.otsuka.co.jp*

(Received September 30, 2004; accepted November 10, 2004; published online December 2, 2004)

Abstract

Statistical analysis was conducted to study the efficiency of methods for active compound selection by comparing the method of screening a test compound set by an *in silico* method DOCK using a target protein (receptor) of a known structure versus the method of screening by the standard *in vitro* assays and also to determine how best to utilize the DOCK output variables. In this study we used DOCK output data on 327 compounds and the *in vitro* assay data on synthetic product resulting from an enzymatic reaction of a given substrate, and those compounds giving greater than 50% inhibition activity in an *in vitro* assay were considered to be active compounds. The representative variables were selected from a group of variables with mutually high correlation in the 108 DOCK output variables and subjected to liberal variable selection or conservative variable selection by the stepwise selection-elimination method of the logistic regression model, yielding 16 and 3 variables, respectively. These variables were then used for screening by the logistic regression method, and the performance was evaluated by the jackknife method (a performance evaluation method in which a measured value predicted from the n-1 observations removing the own predicted observation). The results indicated that elimination of about 80% of test compounds by DOCK *in silico* screening gave 80% sensitivity and 15% false positive rate. We demonstrate the usefulness of *in silico* screening using a prediction model by logistic regression.

Key Words: *in vitro* assay, *in silico* screening, DOCK, logistic regression, variable selection, jackknife method

Area of Interest: Molecular Recognition

1. Introduction

Recent progress in structural biology and genome analysis tools has led to dramatic changes in drug discovery. The drug discovery field has also been influenced by the introduction of combinatorial synthesis technology that permit the synthesis of large amounts of compounds in a short period of time and the high throughput screening (HTS) technology that performs rapid *in vitro* assays [1].

The first requirement of drug discovery is the identification of a compound (active compound) that possesses at last a given level of binding capacity toward a target protein (receptor) among many searchable compounds (test compound set). Traditionally, such a process has utilized *in vitro* assays, but recently, methods have been developed to evaluate systematically the structure and chemical potentials of the target protein and the test compounds through mathematical equations and computer analysis (the *in silico* method). DOCK (version 5.1.0) is an example of such simulation software [2]. Since the *in silico* method is simpler than *in vitro* assays, it can improve efficiencies in drug development, from the standpoint of time, cost, and workload, if such a technology can be applied to drug screening [3].

Thompson et al., [4] have compared the proportion of active compounds in a compound set (hit rate) selected by HTS and the hit rate by an *in vitro* assay conducted after *in silico* screening to pare down the number of compounds and have reported that the hit rate after the *in silico* method is 1700 times the hit rate by HTS alone. However, the study of Thompson et al., compared the HTS versus the *in silico* method using different test compounds and thus cannot be used to determine the generalizability or relevance of this approach. Thus, we compared the direct *in vitro* assay of test compounds and the approach of *in vitro* assay after narrowing down of the same test compounds by *in silico* screening to determine the utility of the *in silico* approach. We also assessed the optimal ways to use the *in silico* method. The methods and results are described below.

2. Method

2.1 Data set

The data used in the study reported here are the results obtained from 327 compounds (test compounds) selected at random from about 30,000 compounds in-house library based either on an *in vitro* assay that measures the product of an enzymatic reaction on a given substrate and the 108 output variables obtained from DOCK. The three-dimensional (3D) structure of the target protein was predicted by a comparative modeling method, which is generally recognized as a highly accurate method under a certain experimental condition [5]. The 327 compounds contained 27 active compound, and the objective of the current study is to keep these 27 compound as much as is feasible, and at the same time by screening attempt to eliminate as many inactive test compounds as possible prior to the *in vitro* assay. In other words, this procedure attempts to eliminate as much as possible the inactive compounds from the test compound set identified by screening.

In general, there is a contradiction between these requirements, and therefore to achieve the former objective of keeping 100% of the active compounds is difficult. Thus, we aimed for the identification of about 80% of the compounds and maximizing the elimination of inactive compounds to achieve good screening and performance characteristics.

In this study, “active compound” refers to compounds showed is greater than 50% inhibition activity by the *in vitro* assay. The *in vitro* assay used here involves a 96 well-plate experiment with three groups, the control group which has the enzyme (target protein), assay buffer and

substrate, the non-treatment group which has the assay buffer and substrate, and the test group which has enzyme, compound, and substrate. After the reaction the measurements made using radioisotope (RI). The compound concentration was 10 μM.

2.2 DOCK output variables

The DOCK software addresses the problem of "docking" molecules to each other. It explores ways in which two molecules, such as a drug and an enzyme or receptor, might fit together. Compounds, which dock to each other well, have the potential to bind.

The DOCK software generates many possible conformations of putative ligand within a user-selected region of a receptor structure. These conformations may be scored using several schemes designed to measure steric and/or chemical complementarity of the receptor-ligand complex. These scores may be used to evaluate likely conformations of a single ligand, or to rank molecules from a library.

The DOCK output variable is an evaluation score that calculates how readily the test compound binds to the target protein by using the given functions. The physicochemical constants needed to determine the evaluation score and the parameters expressing the binding by the compound are included in the DOCK, and options include the ability of the user to specify certain parameters and functions. Changing the parameters as an option can lead to changes in the output variable, and modify the utility of DOCK [6][7][8].

If one assigns certain standard values to multiple parameters available as options, evaluation score for a test compound is produced for each of the variables of the given parameter combination. How to utilize these output variables as well as how to select the variables is the subject of this paper.

In the current study, we determined the standard values to 4 parameters as indicated below, and assessed the 108 variables that result from various combinations of these parameters.

(1) Discovery space

This parameter identifies the binding site on the target protein. The discovery space is defined as the location within the radius R from the active site, and R was assigned to be one of four values, 5, 6, 7, or 8 Å.

(2) Scoring function

This is a parameter that specifies whether the intermolecular interaction between the target protein and the test compound is expressed as energy (E) or contact (C). The evaluation may involve an evaluation in which the test compound is docked to the active site on the target protein (initial evaluation score; primary score) or an evaluation that determines the final interaction score (final evaluation score; secondary score). The primary score and the secondary score can give the four combinations of E/E, E/C, C/E, and C/C. Since C/E can give results that are difficult to interpret, the current study assessed three combinations of E/C, E/E, and C/C.

Based on the instruction manual [2], the energy involves the evaluation of the van der Waals interaction and Coulomb's force, while the contact involves the evaluation of the number of atoms other than hydrogen atoms interacting between the target protein and the test compound.

(3) Distance tolerance

DOCK assumes the shape of the active site to be spherical and searches for the test

compound conformation by placing the atom of the test compound at the sphere center. This parameter can be viewed as the uncertainty in the distance comparisons or sphere centers. Here, three values were assessed, 0.25, 0.5, and 1.0 Å.

(4) Conformation number

This parameter is the number of conformations of the test compound at the active site. Among the conformations restricted by this parameter, DOCK selects the one with the largest evaluation score as the final structure. Here, we studied three values, 25, 100, and 200.

The total number of combinations is 108 ($4 \times 3 \times 3 \times 3$), and since each combination yields a single variable, the 108 variables corresponding to the rows shown in Table 1 are the DOCK output variables studied in this report.

Table 1. Example of parameter combination in output variables.
Each row corresponds to a DOCK output variable.

Discovery space	Scoring function	Distance tolerance	Conformation number
5	E/C	0.25	25
5	E/C	0.25	100
5	E/C	0.25	200
5	E/C	0.5	25
5	E/C	0.5	100
5	E/C	0.5	200
5	E/C	1	25
5	E/C	1	100
5	E/C	1	200
5	E/E	0.25	25
5	E/E	0.25	100
5	E/E	0.25	200
5	E/E	0.5	25
5	E/E	0.5	100
5	E/E	0.5	200
5	E/E	1	25
5	E/E	1	100
5	E/E	1	200
5	C/C	0.25	25
5	C/C	0.25	100
5	C/C	0.25	200
5	C/C	0.5	25
5	C/C	0.5	100
5	C/C	0.5	200
5	C/C	1	25
5	C/C	1	100
5	C/C	1	200
6	.	.	.
6	.	.	.
7	.	.	.
7	.	.	.
8	.	.	.
8	.	.	.

2.3 Model used for screening

As an example, let there be a collection of compounds (test compound population) containing the compound library used here. It may include several active compounds. The positive rate π (population positive rate) depends on the molecular features and chemical characteristics of the

compounds docked to each site, and these molecular features and chemical characteristics are expressed as DOCK output variables. Thus, the positive rate π is a logistic function of the DOCK output variable (x_1, x_2, …, x_r) expressed as equation (1).

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r)} \qquad (1)$$

Under this assumption, β_0, β_1, …, β_r are estimated from data, and for each test compound, DOCK output variable is used to determine the estimated value p for the π , and if p is greater than a given cutoff value, then the test compound is determined to be placed in the selected set, namely the active compound candidate set. This is the screening method proposed in this study.

2.4 Variable selection method in the logistic regression

In this study, the information for the screening considered 108 DOCK output variables, but not all affect the population positive rate. If the small portion of variables have a high degree of predictability, these are sufficient for screening. Thus, the use of equation (1) for screening requires the use of only a portion of 108 variables.

It is thus relevant to consider which variables to should be used. It is appropriate to use the variable selection method as determined by statistical analysis.

That is, the Spearman rank correlation coefficient between variables is calculated, and for a group of variables where the correlation coefficient is mutually 0.9 or greater, one representative variable is taken, while others are discarded.

Then by using the variable selection method in logistic regression analysis [9][10], the variables involved in predicting the positive rate are selected. There are many variable selection methods, and in this study we used the stepwise selection-elimination method with the significance level of 20% or 5%. The former predicts the positive rate with many variables, while the latter predicts the positive rare with a few variables.

The regression coefficient was estimated by the maximum-likelihood method based on standard statistical software. For the actual calculations the SAS® Logistic procedure was used [11].

2.5 Criterion for performance evaluation of the proposed screening method

In screening by determining the prediction value for the population positive rate by the logistic regression model after variable selection, the jackknife method (leave-one-out method [12][13]) was used in this study to identify the variable set best suited for the most efficient selection of active compounds. That is, using the selected variable set, a logistic regression model was created from a test compound set in which one observation was excluded from a test compound set, and the regression equation was used to predict the positive rate of excluding observation. This process is serially repeated by excluding one observation to determine at what proportion (sensitivity) the active compounds are included in the final selected set.

The sensitivity changes depending on the cutoff value. If the cutoff value is small, the sensitivity increases, but the proportion of inactive compound in the selected set (false positive rate), i.e., 1 – specificity, also increases. Thus we specified two types of objectives in performance evaluation.

One is the false positive rate when the sensitivity is at 80%. Of course it is desired that the false positive rate be small.

Another is the area under the curve (AUC), in which the sensitivity with the cutoff value

continuously varied is plotted on the vertical axis and the false positive rate is plotted on the horizontal axis (receiver operating characteristic curve; ROC curve). The ROC curve becomes a curve that connects from the lower left corner to the upper right corner in a square region of the size of 1. In a totally random screening method the curve becomes a straight line connecting the lower left corner (0, 0) and the upper right corner (1, 1); in contrast in a screening method with the sensitivity of 100% and false positive rate of 0%, the curve is a bent curve that passes through the left edge and the upper edge. In the former case AUC = 0.5, and in the latter case AUC = 1, and values closer to 1 indicate better screening performance. The AUC is the evaluation score for the screening method that takes into account various cutoff values.

3. Result

3.1 Correlation between variables

The left figure in Figure 1 shows the correlation coefficients (Spearman rank correlation coefficient) as the heights in 27 DOCK output variables when the discovery space is set at 5 Å. The left-right axis and the front-back axis represent ordering of the 27 variables of the combinations of the scoring function, distance tolerance, and conformation, and from the symmetry of the correlation coefficients, the height of the correlation coefficient is shown only for the lower left half. Since this is a three-dimensional graph, it may appear that the height exceeds 1, but of course none of the values exceed 1. There are a few where the correlation coefficient around 0 becomes a negative value, but these have been left out so that only positive heights are shown in the figure.

The right figure in Figure 1 is a figure of the correlation coefficients in which the discovery space parameter is at 5 Å, resulting in 27 variables (left-right axis), or at 6 Å, resulting in 27 variables (front-back axis).

Only one of the variable pair (the pair of variables on left-right axis and front-back axis) with the height of 0.9 or greater in this figure is kept. Results of analysis of such data indicated that when the scoring function parameter is C/C, the height (correlation coefficient) was 0.9 or greater. From the variables with the scoring function parameter of C/C, one variable was kept while other variables were discarded, and 73 variables were retained for the next step.

3.2 Liberal variable selection

For the 73 variables selected in section 3.1, variable selection was performed by the stepwise selection-elimination method with the significance level of 20%, and the 16 variables shown in Table 2 were selected.

3.3 Conservative variable selection

For the 73 variables selected in section 3.1, stepwise selection-elimination method was applied at the significance level 5%, and 3 variables with the combination of discovery space, scoring function, distance tolerance, and conformation of (5, E/C, 0.5, 100), (6, E/C, 0.5, 200), and (7, E/C, 1, 200) were selected

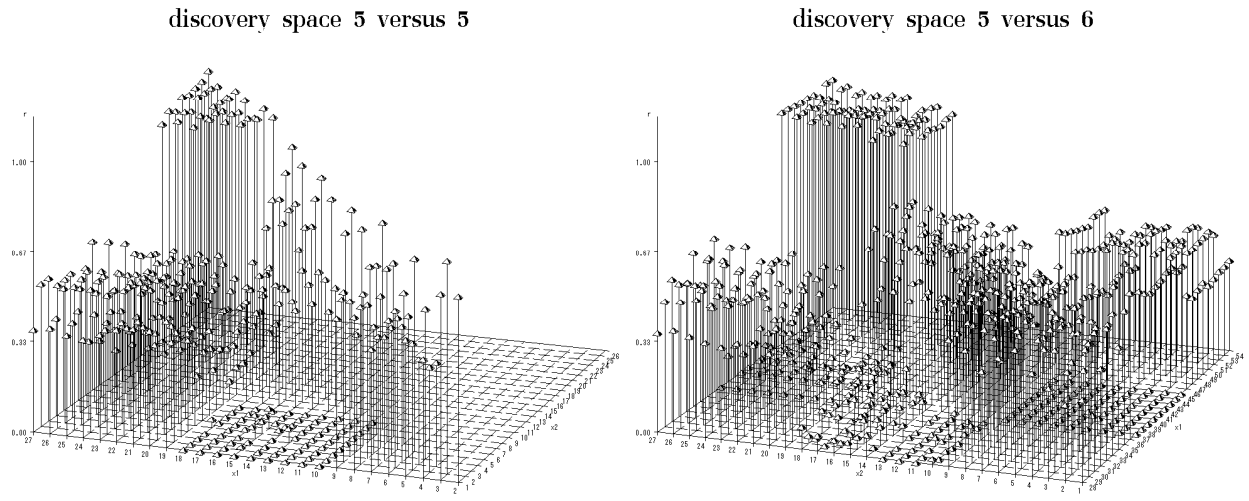

Figure 1. Example of correlation coefficients between variables.
The vertical axis corresponds to correlation coefficient between two variables placed on the two horizontal axes. The same set of variables is placed in both axes in the left figure and different two sets of variables are placed in each axis in the right figure. With respect to two variables with correlation coefficient greater than 0.9, one variable was eliminated from further examination.

Table 2. Selected 16 variables and the estimate of corresponding regression coefficients in logistic regression.

Parameter	Discovery space	Scoring function	Distance tolerance	Conformation number	Estimate	Standard Error
Intercept					-12.556	3.266
x1	5	E/C	0.5	100	-0.027	0.011
x2	5	E/C	1	25	0.012	0.008
x3	5	E/C	1	100	-0.021	0.010
x4	5	E/E	0.25	100	0.002	0.001
x5	6	E/C	0.5	100	-0.018	0.011
x6	6	E/C	0.5	200	-0.028	0.012
x7	6	E/C	1	25	-0.015	0.008
x8	6	E/C	1	200	0.029	0.011
x9	7	E/C	0.5	25	0.015	0.011
x10	7	E/C	0.5	200	0.023	0.009
x11	7	E/C	1	25	0.027	0.010
x12	7	E/C	1	100	-0.051	0.014
x13	7	E/C	1	200	0.046	0.015
x14	7	E/E	0.5	25	0.044	0.019
x15	7	E/E	1	25	0.001	0.000
x16	8	E/E	0.5	200	-0.370	0.118

3.4 ROC curve and AUC

Based on the selection above, the set of 16 variables (first variable set; No.1) and the set of 3 variables (second variable set; No.2) were obtained. For these variable sets the logistic regression equation was determined, and by the jackknife method the sensitivity and false positive rate were determined for various cutoff values to give a ROC curve shown in Figure 2. For reference Figure 2 also shows the ROC curve with a single variable using the best variable (third variable set; No.3). The third variable set is (6, E/C, 1, 25).

Based on Figure 2, at the sensitivity of 80%, the first variable set had a false positive rate of 15%, the second variable set had a false positive rate of 48%, and the third variable set had a false positive rate of 57%. The first variable set had far better performance compared to the other 2 sets.

If the sensitivity is set at 90%, the false positive rates for the three sets are 49%, 53%, and 78%, respectively. Since there are only 27 active compounds among the 327 compounds in the test compound set, sensitivity of 90% requires that 25 active compounds be identified in the selected set, so that the false positive rate is higher overall, and the apparent advantage of the 3 variable sets becomes less apparent.

When compared by the AUC, the AUCs for the first, second and third variable sets are 0.863, 0.751, and 0.691, respectively. It is again evident that the screening is superior when the first variable set is used.

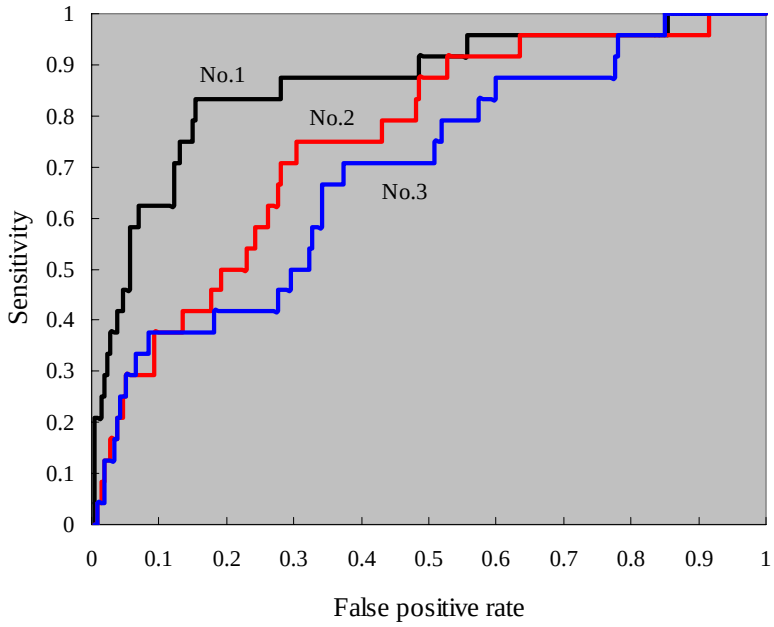

Figure 2. ROC curves for three sets of selected variables.
Horizontal axis: estimated false positive rate, vertical axis: estimated sensitivity.
No.1: 16 variables, No.2: 3 variables, No.3: single variable with the best performance.

4. Discussion

4.1 Role of individual variable

DOCK output variables are used as a set, but there are instances when a variable might be used individually. Analysis of AUC when a single variable is used individually shows that the top 20 variables are as in Table 3. The AUCs tended to be larger when the discovery space is 5 or 6Å, scoring function is E/C, and distance tolerance is 0.5 or 1Å. As a scoring function, E/C tended to be generally better than E/E.

Table 3. The area under the ROC curve in each variable.

Discovery space	Scoring function	Distance tolerance	Conformation number	AUC
6	E/C	1	25	0.691
5	E/C	0.5	100	0.691
6	E/C	0.5	200	0.680
5	E/C	0.5	200	0.673
6	E/C	0.5	100	0.673
6	E/C	0.25	100	0.662
6	E/C	0.25	200	0.661
6	E/C	0.5	25	0.653
7	E/E	0.25	25	0.649
8	E/C	0.5	100	0.642
8	E/C	0.25	200	0.641
7	E/C	1	100	0.636
8	E/C	1	100	0.630
6	E/C	1	100	0.629
5	E/C	1	100	0.626
7	E/C	0.25	200	0.625
8	E/C	0.25	100	0.613
5	E/C	0.25	100	0.609
8	E/E	0.5	200	0.604
8	E/C	1	25	0.601

4.2 Physical meaning of selected set of variables

The reason why the simultaneous use of 16 variables leads better result than the single use of individual variable is probably that each variable reflects different aspect of docking characteristics. Actually, variables in the selected set have physical meanings depicted as follows.

Among the selected 16 variables, there is only one that gives a large discovery space of 8Å. The exclusion of a discovery space of 8Å is interpreted to mean that the discovery space is too large for the size of the active site on the target protein, so that the precision of the binding efficiency decreases, resulting in exclusion of discovery space of this size from the variable selection.

Among the selected 16 variables, there is only one with a small distance tolerance of 0.25Å. The three-dimensional (3D) structure of the target protein used in this study was not derived experimentally from an X-ray crystal structure analysis or NMR but was predicted from structure of

homologous protein determined experimentally by comparative modeling methods, so we interpret this data to indicate that small interatomic distance did not allow sufficient incorporation of the conformational fluctuation of the protein when the test compound binds to the target protein.

Based on these interpretations, we believe that as DOCK parameters, discovery space of 8Å, scoring function of E/E, and distance tolerance of 0.25Å should be excluded as variables.

4.3 Screening rate as a function of sensitivity and false positive rate

By knowing the number of active compounds and test compounds in a given test compound set, it is possible to determine by a simple calculation the screening rate from the sensitivity and the false positive rate, namely the proportion of test compounds remaining in the selected set after screening.

Figure 3 shows a comparison of the 3 variable sets, with the horizontal axis showing the screening rate and the vertical axis showing the sensitivity. With the first variable set, sensitivity of 80% gives a screening rate of 22%. This indicates that if a screening is conducted for the test compounds with π in equation (1) in the upper 22%, 80% of active compound can be found.

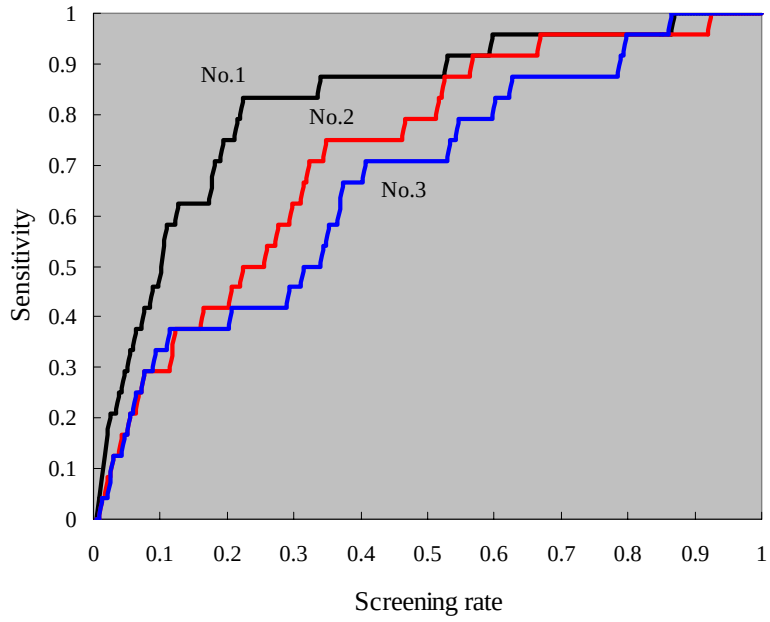

Figure 3. Sensitivity vs screening rate.
The number of compounds in the selected set divided by the number of compounds in the test set is taken on the horizontal axis.
No.1: 16 variables, No.2: 3 variables, No.3: single variable with the best performance.

4.4 Remarks on the actual use of DOCK

Based on the above discussion, use of the set of 16 variables (first variable set) would be the most effective in DOCK. This is of course more efficient than conducting *in vitro* assays directly,

but when one also considers the time and labor involved in screening, it is also necessary to consider simpler methods. While a computer performs the analysis, simulations are conducted, so that the calculation time is dependent on the number of variables, and the time needed for such calculations is not insignificant.

The best method to reduce the time needed is to use the 3 variable set (second variable set) or the 1 variable set (third variable set) as the variable set. The decreased performance in such cases might be best assessed as the screening rate when the sensitivity is 80%. One can then compare the time and cost needed for *in silico* screening versus the time and cost needed for *in vitro* assay. We have not done these calculations, but feel that in actual use situations, such calculations are needed.

In this study, we evaluated performance of an *in silico* screening method using our own compound library and experimental data. The results of this study do not necessarily imply that our variable sets and regression equations can be used for general applications. Future work is needed to determine how applicable the current data are to general situations. At the current state of DOCK applications, it would be necessary to conduct for example *in vitro* assays using some appropriate number of compounds, for example 100 compounds, and determine the best variable set and regression formula by the methods described here.

It is emphasized that screening with 1 variable with given option parameters is extremely inefficient.

5. Conclusion

This study showed that based on our 327 compound data, *in silico* screening with good performance is feasible.

In the field of drug discovery, HTS is used to react compounds with the target protein at a high speed using robotics, but if one estimates that the cost of the *in vitro* assay per test compound is about 100 yen, then testing 1 million or more test compounds would cost 100 million yen [14]. If the screening by the method described in this study can decrease the number of test compounds to be screened by 50%, then the costs decrease by one half.

To develop a completely new drug requires about ten years and cost of 50 billion yen [15]. The results of the current study show that in the early phase of the drug discovery process, one can omit the time-consuming *in vitro* assays and narrow down the number of active compounds to be developed. This approach may shorten the development timelines and decrease development costs.

The authors express their sincere thanks to Doctor Hiromi Uchiro of Tokyo University of Science for his advice and suggestions that were helpful in the preparation of this paper. The authors also express their sincere thanks to the two referees for their fruitful advices on the revision of our manuscript.

This research was partially supported by JSPS Grant-in-Aid for Scientific Research No. 16200022.

References

[1] T. Kume, T. Harada, K. Fukuda and H. Shimadzu., *Drug Metab. Pharmacokin. in Japanese*, **16**, 162-163 (2001).
[2] Ewing T., Makino S., Skillman A. and Kuntz I., *J. Comput. Aided. Mol. Des.*, **15**, 411-428 (2001).
[3] Y. Matsuo, *Medical Science Digest in Japanese*, **28**, 193-196 (2002).
[4] Thompson N. Doman, Susan L. McGovern, Bryan J. Witherbee, Thomas P. Kasten, et al., *J. Med. Chem.*, **45**, 2213-2221 (2002).
[5] David Baker and Andrej Sali, *SCIENCE*, **294**, 93-96 (2001).
[6] Ewing T. and Kuntz I., *J. Comput. Chem.*, **18**, 1175-1189 (1997).
[7] Ryan T. Koehler and Hugo O. Villar, *J. Comput. Aided. Mol. Des.*, **14**, 23-37 (2000).
[8] C.M. Oshiro, I.D. Kuntz and J. Scott Dixon, *J. Comput. Aided. Mol. Des.*, **9**, 113-130 (1995).
[9] Miller AJ., Subset selection in regression, Chapman and Hall (1974).
[10] Allen DM., *Technometrics*, **16**, 125-127 (1974).
[11] SAS Institute Inc., SAS Technical Report J-119 LOGISTIC Procedure, SAS Institute Inc. (1994).
[12] Miller RG., *Biometrika*, **61**, 1-15 (1974).
[13] C. Hamada, SAS Users Group International Japan 2000 Proceedings, 13-38 (2000).
[14] Burbaum JJ, *Drug Discov. Today*, **3**, 313-322 (1998).
[15] Jurgen Drews and Stefan Ryser, *Drug Discov. Today*, **2**, 365-372 (1997).