English translated document of "Journal of the Institute of Science and Engineering. Chuo University Vol. 25 2019"

In silico screening based on the protein–ligand interaction of experimental data

Takaaki ICHIKAWA*, Hideaki UMEYAMA**, Mitsuo IWADATE**

Abstract

The prediction accuracy of ChooseLD, a technique for *in silico* screening based on a three-dimensional structure, was remarkably improved by introducing hydrophobic interactions to prediction parameter.

Specifically, the Hc index, which optimized the coefficient, for the evaluation of hydrophobic interactions, was included in formula. The hydrophobic interaction was found to have a remarkable effect as a physical parameter. The effectiveness of the combination of an empirical parameter and the physical parameter is demonstrated simultaneously.

1. Introduction

In recent years, an increasing number of interactions between drugs and proteins have been identified through biochemical experiments [1][2]. Consequently, the competition to discover compounds that inhibit target proteins has intensified among pharmaceutical companies and research institutions [3][4]. Numerous compounds that inhibit target proteins have been found through experimental screening. However, owing to the high cost needed for the experimental screening, *in silico* screening using computers is preferable.

In *in silico* screening, the three-dimensional arrangement of target proteins can be determined by utilizing experimental structures deposited in the Protein Data Bank (PDB) [5], such as those obtained through X-ray crystallography and NMR. Alternatively, multiple aqueous solution structures are optimized through energy minimization and molecular dynamics calculations to derive the atomic coordinates of the target protein. In contrast, if the three-dimensional structure is not registered in database, it is necessary to predict the three-dimensional structure using protein structure prediction methods, such as homology modeling [6]. In either case, data obtained through experimental protein structure analysis or prediction are used.

Additionally, the number of protein–ligand complex structures registered in the PDB has increased in recent years. It is also common to find multiple X-ray structures within a single protein family, each containing ligands comprised of different atoms [1][2]. Several attempts are being made to evaluate whether the complex structures predicted by docking software align with these data.

* Graduate School of Science and Engineering, Major in Life Sciences, Chuo University
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
** Department of Life Sciences, Faculty of Science and Engineering, Chuo University
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan

In these attempts, existing docking software was used to dock candidate inhibitor compounds to target proteins from a virtual compound library for predicting the structures of protein–ligand complexes. After predicting structures, inhibition activities of compound are tested to select additional hit compounds. Additionally, calculating distances between the protein and ligand, classical physical energies, and extracting and re-evaluating interaction information from the structures of known protein–ligand complexes are assessed [7][8].

The ChooseLD (CHOOse biological information Semi-Empirically on the Ligand Docking) [9], method used in the present study, can efficiently select key information from the data of known protein–ligand complexes structure registered in the PDB and perform docking. It is also useful for detecting a large number of hit compounds. However, when performing homology modeling based protein structure prediction, docking results could be mis-calculated since citing the information of the proteins with mutations in the ligand-binding site or proteins sharing structural similarities, although that are strongly denied by hydrophobic interactions influence.

In this study, to improve the accuracy of the ChooseLD for predicting protein–ligand complexes interactions, we introduced the hydrophobic interaction evaluation function Hc (hydrophobic correlation) index [10], conceived by Hideaki Umeyama and Kenji Akaba to score calculation.

In recent years, there has been a notable progress in the development and release of large-scale databases useful for drug discovery research. One such database is ChEMBL [11], which provides structure–activity relationship (SAR) information for pharmaceuticals and developmental compounds. High-quality SAR information, collected and organized by chemistry and biology experts, is provided free of charge. By correlating the activity information (Ki) of targets registered in ChEMBL, we investigated whether integrating the hydrophobic interaction evaluation function Hc index into the ChooseLD method could enhance the efficacy in exploring compound candidates as opposed to the original ChooseLD method.

2. Methods

2-1. Outline

ChooseLD is a software tool that extracts biochemical information from protein–ligand complexes with known interactions registered in the PDB and performs docking simulation. It utilizes fingerprint-based docking of unfamiliar ligand structures, incorporating partial binding free-energy components derived from ligands with known interactions.

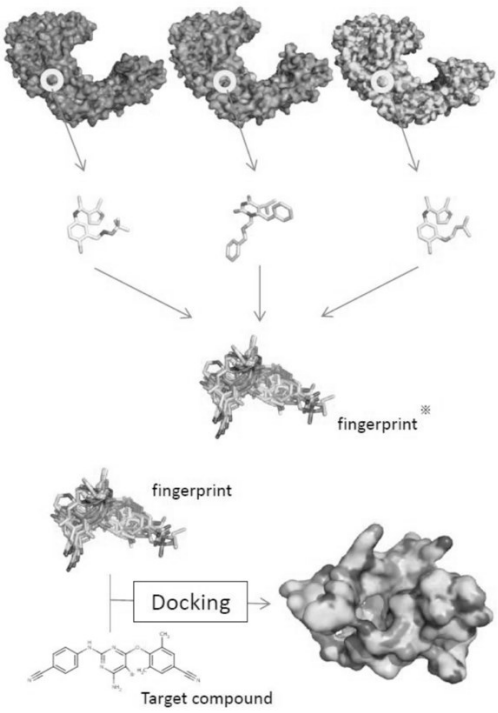

Fig. 1 Flow of docking simulation using the ChooseLD method.

An overview of the ChooseLD docking simulation is shown in Fig. 1. The ChooseLD method utilizes the structural information of similar protein for target protein and identifies ligands that bind to similar sites on these proteins. This process generates fingerprints, which are components retaining partial binding free energy from ligands with known interactions. Compounds are docked based on this fingerprint information. By performing docking based on ligand information from proteins with similar local structures to the target protein, it is possible to explore potential pharmaceutical compounds that act on the target protein.

However, if mutations occur in the referenced local structures or in the ligand-binding site of the target protein, docking a hydrophilic ligand to a hydrophobic binding site of the target protein could lead to inaccurately high evaluations of the protein–ligand complex structures. Conversely, the same issue can occur when attempting to dock hydrophobic ligands onto hydrophilic binding sites. To address this, the Hc index was incorporated.

2-2. Hf index

The Hc index was originally created as an indicator for assessing the status of hydrophobic and hydrophilic interactions between proteins and compounds, and it is derived from the Hf (hydrophobic field–effect) index. The Hf index is explained first.

The Hf index is an indicator representing the strength of hydrophobic and hydrophilic interactions that a compound receives from a protein. To calculate the Hf index, the surface of

the compound is first divided into micro-surfaces (patches), and the calculation is performed for each patch using equation (1):

$$Hf_j = \sum_k f_k \cdot \varphi \qquad (1)$$

$$\varphi = \exp\left\{\beta \cdot \left(r_{jk} - R\right)^2\right\} \qquad (2)$$

Hfj represents the Hf index in the j-th patch. f_k is the transfer free energy per unit area of the k-th atom of the protein. Here, transfer free energy refers to the change in free energy when a molecule moves from the aqueous phase to the gas phase. This transfer free energy is closely associated with the hydrophobic and hydrophilic nature of the molecule, as already described by Ben-Naim [12]. The f value, which quantifies transfer free energy, indicates strong hydrophilicity when it takes positive values and strong hydrophobicity when it takes negative values. These values have already been determined at the functional group level by previous research. [10]

Table. 1 The f value for the protein functional group

Substituents	f(cal/(mol · Å^2))
Guanidinium	19.30
-SH	-24.10
-S-	0.71
Imidazolium	1.27
Indolyl	-12.56
$-NH_3^+$	45.28
$-C_6H_5$	-12.88
$-CONH_2$	11.30
$-COO^-$	18.63
-OH(aliphatic)	11.26
-OH(aromatic)	15.78
Hydrocarbon	-20.87
Backbone amide	29.34

ϕ is a distance -dependent function that accounts for the attenuation of the hydrophobic influence of atom k with distance. The method of calculating ϕ is shown in equation (2).

rjk is the distance between the center of patch j and the surface of protein atom k. Here, β is a constant, -0.1312, which is taken from the Pratt-Chandler theory to reproduce a quasi-stable state with one water molecule between two molecules [13]. R is a value taken to reproduce the stable state in the protein [13]. R represents the radius of the atom in the protein, and ($rjk-R$) indicates the distance between the surfaces of the compound and the protein atoms. Thus, the Hf index represents the influence received by the patch from the protein by summing the hydrophobic and hydrophilic strengths of each protein atom, weighted by distance.

2-3. Hc Index

The Hc index can be obtained with following equation (3), using the Hf index:

$$H_c = \sum_j f_j \cdot Hf_j \cdot S_j \qquad (3)$$

S_j represents the area of patch j. Hence, the Hc index is the sum of the hydrophobic and hydrophilic influences (Hfj) each patch receives from the protein multiplied by the f value of the patch itself and the area S_j of the patch. This enables the calculation of the overall hydrophobic and hydrophilic correlation between the compound and the protein. However, although the f values of compounds have been determined by previous research, the amount of data available for practical use in *in silico* screening is limited. Therefore, in this study, we developed a method to determine the f value of the compound patches themselves.

2-4. f value

Figure 2 shows the correlation between the partition coefficient logP in the water-n-octanol system and the transfer free energy f value per unit area.

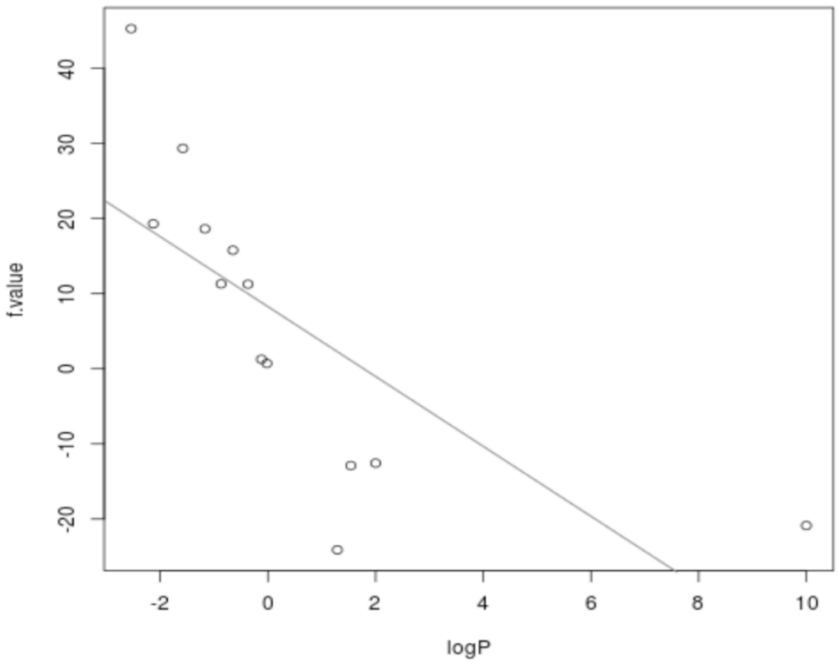

Fig. 2 – Partition coefficient logP and transfer free energy f value per unit area

Because of the negative correlation between logP and f value, a negative sign was attributed to the partition coefficient logP, and it was set as the f value of each compound atom.

$$f_j = -\log P \qquad (4)$$

Additionally, additivity of logP was assumed and, thus, it was determined using the Crippen's method [14]. This was used as the logP per compound atom. In the Crippen's method, the pattern of adjacent atoms to the atom for which logP is to be calculated is matched against predefined patterns in the SMARTS format. The logP of the matching pattern is used as the logP of each atom. An example is shown below.

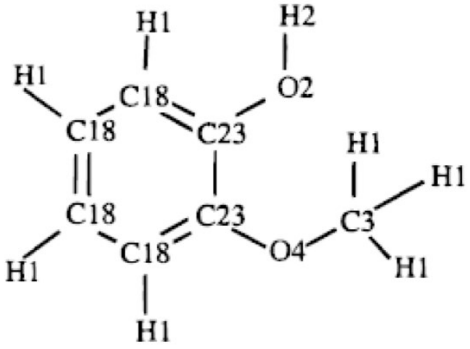

Table. 2 – logP per compound atom

Type	SMARTS	logP
C3	'[CH3][(N,O,P,S,F,Cl,Br,I)]'	-0.2035
C18	'[cH]'	0.1581
C23	'[c](:a)(:a)-O'	0.5437
H1	'[#1][#6]', '[#1][#1]'	0.1230
H2	'[#1]O[CX4]', '[#1]Oc', '[#1]O[!(C,N,O,S)]', '[#1][!C,N,O)]'	-0.2677
O2	'[OH]', '[OH2]'	-0.2893
O4	'[O](A)a', '[O](a)a'	-0.4195

In the compound shown above Table 2, even for carbon atoms that form the same aromatic ring, the carbon atom bonded to the hydroxyl group (type: C23) has a higher logP. Thus, the logP per atom was determined by the adjacent atoms, and the value is multiplied by a sign to calculate the f value of the compound. By incorporating this method, the final Hc Index can be expressed as equation (5) below:

$$H_c = \sum_j -\log P \cdot Hf_j \cdot S_j \qquad (5)$$

2-5. Determining the criterion value of coefficient k

Since the systems differ between the FPAScore calculated by the ChooseLD method and the Hc index, the final score was obtained as the Hc index multiplied by the coefficient k added to the FPAScore, result in including hydrophobic interaction evaluation function Hc index. The equation is shown below (Equation (6)):

$$\text{Total Score} = \text{FPAScore} + k \times \text{Hc Index} \qquad (6)$$

To determine the value that serves as the criterion for coefficient k, all target sequences registered in ChEMBL (ChEMBL20) were used to construct three-dimensional structures and an FP Library. Additionally, targets with a sufficient number of compounds in the same assay were identified. Next, ChooseLD was applied to these targets to obtain FPAScores, and the structure with the highest FPAScore output by ChooseLD was identified. The Hc index for this structure was then calculated. Finally, the variance of each of these scores was calculated, and the ratio of these variances was used as the criterion value for the coefficient k.

The FPAScore was found to be 565.5944 ± 415.2728, and the Hc index was 0.3959617 ± 12.76401. Owing to the significant difference in the range of variances, we centered around k = 32.5 and varied the value of k, correlating it with experimental values (Ki) registered in ChEMBL (ChEMBL20). Through this process, the optimal coefficient k was determined.

3. **Results and discussion**

3-1. Determining the optimal coefficient k

Out of the 5,569 targets registered in ChEMBL (ChEMBL20), 4,348 targets were successfully constructed using LigandFAMS. Among these, 596 targets had the necessary assay information and Ki data required for correlating the calculated scores with experimental values. Of these, ChooseLD was successfully executed on 273 targets. The information obtained for each target was compiled according to the assay ID, as experimental conditions and values such as Ki can differ among assays even for the same compound.

To verify the improvement in accuracy, two types of correlations were conducted: one between Ki and the scores calculated by the original ChooseLD program, and the other between Ki and the scores calculated by the ChooseLD program including the Hc index. In the analysis of these correlation, Spearman's rank correlation was used.

To compare the correlation between Ki and the original ChooseLD scores, and the correlation between Ki and the scores including the Hc index, a null hypothesis stating "there is no difference in the rank correlation coefficient ρ" was formulated, and a t-test was performed. The results of these tests for each k are shown in Table 3:

.

Table 3. Test results of correlation analysis for each k

k	Two-tailed test	One-tailed test (less)	One-tailed test (greater)
27.5	0.5078	0.7461	0.2539
28.5	0.2288	0.8856	0.1144
29.5	0.463	0.7685	0.2315
30.5	0.0285	0.9858	0.0143
31.5	0.1175	0.9413	0.0588
32.5	0.2554	0.8723	0.1277
33.5	0.0893	0.9553	0.0447
34.5	0.1765	0.9117	0.0883
35.5	0.1602	0.9199	0.0801
36.5	0.2673	0.8663	0.1337

Based on the test results shown in Table 3, the optimal k was determined as 30.5. The correlation coefficient from the original ChooseLD is denoted as “original,” and the correlation coefficient from the improved ChooseLD, which includes the hydrophobic interaction evaluation function Hc index, is denoted as “improved.” The test results for k = 30.5 are explained as follows:

① Two-tailed test
The null hypothesis was set as “there is no difference between the original and improved correlation coefficients.” The test yielded the following result:

p-value = 0.02849

This low value indicates that it is reasonable to reject the null hypothesis. Therefore, it is demonstrated that there is a significant difference between the original and improved correlation coefficient.

② One-tailed test
A one-tailed test was conducted to compare the magnitude of the values between the original and improved ChooseLD.

1. Null Hypothesis: “The original is not smaller than the improved.” The result:

   p-value = 0.9858

   This very high p-value indicates that the null hypothesis is accepted with a probability of 0.9858.

   Null Hypothesis: “The original is not larger than the improved.” The result:

   p-value = 0.01425, a very low value.

Based on these results, it is concluded that including the hydrophobic interaction evaluation function Hc index in ChooseLD is most effective in improving prediction accuracy when k = 30.5.

4. **Conclusions**

An improvement in the predictive accuracy of the *in silico* screening method ChooseLD, which is based on three-dimensional structures, has been observed by taking into account hydrophobic interactions. Specifically, the hydrophobic interaction evaluation function Hc index was included and optimized for the coefficient. The results demonstrated that hydrophobic interactions play a crucial role as a physical parameter. The effectiveness of combining empirical parameters with physical parameters was also comprehensively demonstrated.

5. **References**

［1］ Wood ER, Truesdale AT, McDonald OB, Yuan D, Hassell A, Dickerson SH, A Unique Structure for Epidermal Growth Factor Receptor Bound to GW572016 (Lapatinib): Relationships among Protein Conformation, Inhibitor Off-Rate, and Receptor Activity in Tumor Cells, Cancer Res., 64, 6652-6659 (2004).

［2］ Stamos J, Sliwkowski MX, Eigenbrot C, Structure of the Epidermal Growth Factor Receptor Kinase Domain Alone and in Complex with a 4-Anilinoquinazoline Inhibitor, J. Biol. Chem., 277, 46265-46272 (2002).

［3］ Nakamura K, Yamamoto A, Kamishohara M, Takahashi K, Taguchi E, Miura T, Kubo K, Shibuya M, Isoe T, KRN633: A selective inhibitor of vascularendothelial growth factor receptor-2 tyrosine kinase that suppresses tumor angiogenesis and growth, Mol. Cancer. Ther., 3, 1639-1649 (2004).

［4］ Nakamura K, Taguchi E, Miura T, Yamamoto A, Takahashi K, Bichat F, Guilbaud N, Hasegawa K, Kubo K, Fujiwara Y, Suzuki R, Kubo K, Shibuya M, Isoe T, KRN951, a Highly Potent Inhibitor of Vascular Endothelial Growth Factor Receptor Tyrosine Kinases, Has Antitumor Activities and Affects Functional Vascular Properties, Cancer Res. 66, 9134-9142 (2006).

［5］ Umeyama H, Iwadate M, FAMS and FAMSBASE for protein structure, Curr. Protoc. Bioinformatics, Chapter 5: Unit 5.2 (2004).

［6］ Westbrook J, Feng Z, Chen L, Yang H, Berman HM, The Protein Data Bank and structural genomics, Nucleic Acids Res. 31, 489-491 (2003).

［7］ Muralia S, Hojob S, Tsujishitaa H, Nakamurac H, Fukunishi Y, In-silico drug screening method based on the protein-compound affinity matrix using the factor selection technique, Eur. J. of Med. Chem., 42, 966-976 (2007).

［8］ Deng Z, Chuaqui C, Singh J, Structural Interaction Fingerprint (SIFt): A 9Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions, J. Med. Chem. 47, 337-344 (2004).

［9］ Takaya D, Takeda-Shitaka M, Terashi G, Kanou K, Iwadate M, Umeyama H, Bioinformatics based Ligand-Docking and in-silico screening, Chem. Pharm. Bull., 56, 742-744 (2008).

［10］ Kenji Akahane, Yasuo Nagano, and Hideaki Umeyama, Hydrophobic Effect on the Protein-Ligand Interaction; Hydrophobic Field-Effect Index and Hydrophobic Correlation Index, Chem. Pharm. Bull., 37, 86-92 (1989).

［11］ Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP, ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic Acids Res., 40, D1100-D1107 (2012).

［12］ Ben-Naim A, Statistical Mechanical Study of Hydrophobic Interaction. III. Generalization and Further Applications, J. Chem. Phys., 57, 5257-5265 (1972).

［13］ Pratt LR, Chandler D, Theory of the Hydrophobic Effect, J. Chem. Phys., 67, 3683-3704 (1977).

［14］ Scott A. Wildman and Gordon M. Crippen, Prediction of Physicochemical Parameters by Atomic Contributions, J. Chem. Inf. Comput. Sci., 39, 868-873 (1999).