International Journal of High Throughput Screening

Dovepress

open access to scientific and medical research

Open Access Full Text Article

METHODOLOGY

Quantitative analysis of aggregation-solubility relationship by in-silico solubility prediction

This article was published in the following Dove Press journal:
International Journal of High Throughput Screening
25 June 2010
Number of times this article has been viewed

Tadaaki Mashimo[1,2]
Yoshifumi Fukunishi[3]
Masaya Orita[2,4]
Naoko Katayama[2,4]
Shigeo Fujita[2,5]
Haruki Nakamura[3,6]

[1]Information and Mathematical Science Laboratory Inc., Tokyo, Japan; [2]Japan Biological Informatics Consortium (JBIC), Tokyo, Japan; [3]Biomedicinal Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan; [4]Chemistry Research Labs, Drug Discovery Research, Astellas Pharma Inc., Ibaraki, Japan; [5]Astellas Research Technology Inc., Ibaraki, Japan; [6]Institute for Protein Research, Osaka University, Osaka, Japan

Abstract: Aggregator (frequent hitter) compounds show non-selective binding activity against any target protein and must be removed from the compound library to reduce false positives in drug screening. A previous study suggested that aggregators show high hydrophobicity. The LogS values of aggregators and non-aggregators were estimated by the artificial neural network (ANN) model, the multi-linear regression (MLR) model, and the partial least squares regression (PLS) models, with the weighted learning (WL) method, and the results showed the same trend. The WL method is weighted on the data of the learning set molecules that are similar to the test molecule and improves the prediction accuracy. Bayesian analysis was applied, revealing a simple relationship between aggregation and solubility. Namely, the molecules with LogS > -5 were non-aggregators. In contrast, most of the molecules with LogS < -5 were aggregators. We also made a simple look-up table of probability of aggregation depending on the molecular weight and the number of hetero-atoms.

Keywords: aggregator, frequent hitter, compound library, solubility prediction, generalized-Born accessible-surface area, GBSA

Introduction

Non-specific compounds are frequently observed in high-throughput screening (HTS). These compounds are called frequent hitters or aggregators. Jadhav et al reported that almost 90% appeared to be detergent-sensitive hits or aggregators of the compounds showing concentration-dependent inhibition in the detergent-free screening assay.[1] The mechanism underlying such non-selectivity is complicated: some aggregators form micelle colloids, while others show nonspecific affinities with many different kinds of proteins. There have been several reports about aggregators, and some aggregator prediction methods have been described.[1–9] In these reports, the methods used were the support vector machine, decision tree, Bayesian model, etc, based on the set of molecular descriptors.[6–9] These methods succeeded in the prediction of aggregators and several features of aggregators have been reported: rigidity, hydrophobicity, numerous aromatic rings, and so on. However, these features are also common in known drugs. The quantitative relationships between aggregation and physical properties have remained unclear. Therefore, we investigated the relationships between aggregation and aqueous solubility (LogS) by using our LogS prediction method.

There have been many reports published about the logS prediction methods.[10–27] The most popular methods for predicting LogS involve a type of regression (multi-linear regression (MLR), partial least squares (PLS) regression) based on the molecular descriptors, which represent the atoms and substructures (group contribution method)

Correspondence: Yoshifumi Fukunishi
Biomedicinal Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-3-26, Aomi, Koto-ku, Tokyo 135-0064, Japan
Tel +81 3 3599 8290
Fax +81 3 3599 8099
Email y-fukunishi@aist.go.jp

within the molecule. Some methods consider several physical properties, such as surface area and molecular volume, as descriptors.[16,18]

Solubility can be approximated by experimentally observed properties, such as melting point (MP) and LogP.[17]

$$\text{LogS} = 0.5 - 0.01(\text{MP} - 25) - \text{LogP} \quad (1)$$

The MP (°C) and LogP are easily observed experimentally, in contrast to the LogS. The MP represents the stability of the crystal, and the transfer free energy from pure solvent to octanol is assumed to be constant and independent of the solute. This method showed that solubility can be analyzed as the combination of several processes of solvation, which is a simple physical process. Jorgensen and Duffy showed that the LogS value could be predicted by the solute-solvent Coulombic interaction, the van der Waals interaction, the accessible surface area and the numbers of several kinds of functional groups included in the molecule.[16] The LogS value could be approximated by only these six terms, and the predicted value showed a quite high correlation to the experimental data; the coefficient of determination (R^2) was 0.9.

To examine the relationship between LogS and aggregation, we improved the LogS estimation method. We introduced a weighted learning approach, which is weighted on the data of the learning set molecules that are similar to the test molecule. The similarity between molecules is calculated based on a set of molecular descriptors. The descriptors contain some physical properties as solvation free energy and accessible surface area, considering the previous work described above.[16] The LogS values of aggregators and non-aggregators were estimated by applying the WL method to the MLR, ANN, and PLS models. These three methods suggested the same clear correlation between LogS and aggregation; that is, the molecules with LogS $>$ −5 were non-aggregators. In contrast, most of the molecules with LogS $<$ −5 were aggregators.

Methods

We applied three LogS predictors: MLR, ANN, and PLS. The definitions of MLR and PLS are clear, so we have omitted an explanation of these two methods. Our ANN LogS predictor uses an error-back propagation method with a set of molecular descriptors[28] and it was developed using the MolWorks® software framework (Beyond Computing Co. Ltd., Tsukuba, Japan).[29]

To improve the prediction accuracy, we introduced the weighted learning method with these three predictors. Without the weighted learning (WL) method, each molecule in the learning set is learned once. With the WL method, each molecule in the learning set is learned several times, depending on the similarity to the test (query) molecule. A molecule that is similar to the test molecule is learned several times, while a molecule that is not similar to the test molecule is learned once. The number of learning procedures depends on the similarity to the test molecule.

The relationship between the aggregation and the solubility is analyzed by the Bayesian statistics. The probability of aggregation is approximated by a sigmoid function of LogS value. This model estimates the probability of aggregation of subset of compound library.

Descriptors

We developed a new descriptor by modifying the MolWorks software.[30] Table 1 shows the set of the molecular descriptors used in the current study. Newly added descriptors in the current study are the 2nd–17th descriptors in Table 1. The LogS value could be approximated by the melting point and the LogP value, as shown in equation (1) above. The LogP value of a compound was calculated from the transfer free energy of the compound from octanol to water, as evaluated by the generalized-Born accessible-surface area (GBSA) method.[31,32] This method can estimate the transfer free energy from a vacuum to a solvent with a specific surface tension and dielectric constant. The molecular structure (dominant ion form) can change in the solvation process. In the current study, the dominant ion form of the COOH group in water is COO^-. To simplify the problem, only two molecular structures were prepared for each molecule, when possible. Two ion forms were prepared for carboxylic acid, sulfuric acid, phosphoric acid, and amines. Namely, $-COO^-$ for –COOH, $-SO_3^-$ for $-SO_3H$, $-PO_3^{2-}$ for $-PO_3H_2$, $-PO_2^-$– for $-PO_2H$–, $-NH_3^+$ for $-NH_2$, $-NH_2^+$– for –NH– (secondary amine), and $-NH^+$ $<$ for –N $<$ (tertiary amine). Only one molecular structure was prepared for a molecule without these functional groups. We assumed that the H atom of the C-OH group does not dissociate, and that the dominant ion form of p-nitrophenolate ($C_6H_4NO_3^-$; the CAS No. is 14609-74-6) is an anion. The solvation free energies of these two ion forms into water and octanol were calculated by the GBSA method, and these four energies were adopted as the descriptors. The number of dissociated H atoms that bind to O or N atoms was also adopted as a descriptor.

The other descriptors were the Joback-like descriptors those are the numbers of substructures. In addition, some

Table 1 Molecular descriptors used in the current study. 3: DelO_H: number of dissociated H atoms in water. 4: addN_H: number of added H atoms in water. Charge (–): number of atoms with atomic charges <–0.3. Charge (medium): number of atoms with atomic charges between >–0.3 and <0.3. Charge (+): number of atoms with atomic charges >0.3. dASA_o: accessible surface area per unit molecular volume in octanol. dG_o: transfer free energy from vacuum to octanol. dASA_wd: accessible surface area per unit volume of dissociated molecule in water. dG_wd: transfer free energy of dissociated molecule from vacuum to water. dASA_w: accessible surface area per unit volume of molecule in water. dG_w: transfer free energy of molecule from vacuum to water. 59: $–NH_2$ or $–NH_3$ group that binds the atom with p electrons. 60: –NH– group that binds the atom with p electrons. 61: –NH– group in ring that binds the atom with p electrons. 62: >N– group that binds the atom with p electrons. 63: >N– group in ring that binds the atom with p electrons

1	2	3	4	5
Mass weight	No of aromatic atoms	DelO_H	addN_H	No of atoms in ring
6	7	8	9	10
charge (–)	charge (medium)	charge (+)	dASA_o	dG_o
11	12	13	14	15
dASA_wd	dG_wd	dASA_w	dG_w	ddG_o
16	17	18	19	20
ddG_wd	ddG_w	$–CH_3$, CH_4	$>CH_2$, $>CH_2$ (ring)	>CH–, >CH– (ring)
21	22	23	24	25
>C<, >C< (ring)	$=CH_2$	=CH–	=C<	=C=
26	27	28	29	30
≡CH	≡C–	=CH– (ring)	=C< (ring)	Fluorine
31	32	33	34	35
Chlorine	Bromine	Iodide	Alcohol R–OH	Phenol Ar-OH
36	37	38	39	40
–O–	–O– (ring)	>C=O	>C=O (ring)	–CHO (aldehyde)
41	42	43	44	45
–COOH (acid)	–COO (ester)	=O except (COO, SO_2...)	$–NH_2$, NH_3	>NH, >NH (ring)
46	47	48	49	50
>N–, >N– (ring)	=N–	–N= (ring)	–CN	$–NO_2$
51	52	53	54	55
–SH	–S–	–S– (ring)	>PS–	>PO–
56	57	58	59	60
–(C=S)–, –(C=S)– (ring)	–(S=O)–	–(O=S=O)–, –(O=S=O)– (ring)	$–NH_2$, $–NH_3$ (connected to atom with π orbital)	–NH– (connected to atom with π orbital)
61	62	63		
Ring –NH– ring (connected to atom with π orbital)	>N– (connected to atom with π orbital)	Ring >N– ring (connected to atom with π orbital)		

substructures were added in the current study, since the Joback descriptor does not include P, S, and halogen atoms.

Weighted learning (WL) method

In the WL method, the molecules in the learning set that are similar to the test molecule are learned many times. In contrast, the molecules in the learning set that are not similar to the test molecule are learned only once. The similarity between two molecules is given by a distance, and the distance between molecule A and molecule B (D(A, B)) is a generalized Euclidean distance of the descriptors, defined as:

$$D(A,B)=\sum_{i=1}^{N} c_i(d_i(A)-d_i(B))^2 \qquad (2)$$

where $d_i(X)$ is the i-th descriptor of molecule X, N is the total number of descriptors, and c_i is a normalization coefficient. The value of c_i is defined as the deviation of $d_i = 1$.

Figure 1 shows the schematic representation of how to determine the learning time. For a test molecule X, the distribution of D is calculated, with s as the deviation of D. Let M and h be the number of bins and the bin size, as follows:

$$h = s/M \qquad (3)$$

Let m and L_{max} be integer and the maximum number of learning times. For molecule A in the learning set, if $mh < D(A, X) < (m + 1)h$, then molecule A is learned $L_{max}–m$ times. If $L_{max}–m < 1$, then molecule A is learned once.

In the framework of the ANN and MLR methods, the same data cannot be learned more than once. Thus, the 14th to 17th descriptors in Table 1 are randomly modulated up to 3% to generate the L data for L-times learning. Since molecular structures are flexible and thus the physical properties can change, 3% modulation should be reasonable.

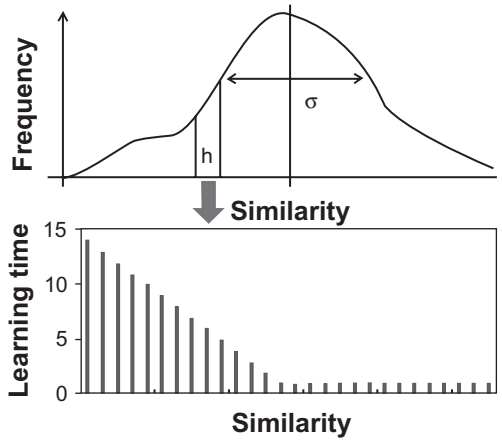

Figure 1 Schematic representation of the relationship between molecular similarity and learning time.

Computational procedures

The preparation of the molecular descriptors consists of several steps:

Step 1: The Joback descriptors were calculated from the molecular structure. The original compound data were described in the SMILES format, and the 3D structures were generated by Molecular Operating Environment (MOE, Chemical Computing Group, Montreal, Canada). Some structures of compounds were modified by visual inspection. The Joback descriptors do not depend on the protonation state of the molecule.

Step 2: Two protonated structures (ion forms) were prepared. One was a dominant ion form in water and the other was a dominant ion form in octanol. These ion forms were prepared by the Hgene software of myPresto® (http://medals.jp/econtents/download and http://presto.protein.osaka-u.ac.jp/myPresto4/index_e.html).

Step 3: For each ion form, several conformers were generated by the conformer generation program confgene of myPresto. The rotatable bonds were randomly rotated in 60 degree increments.

Step 4: The coordinates of each conformer were energy-optimized by using the GBSA model of cosgene/myPresto. The dielectric constants of water and octanol were set to 78.5 and 4.0, respectively. The atomic solvation parameters (surface tensions) of all atoms were set to 10 cal/mol/Å^2. The most stable structures were adopted for water and octanol. The physical properties, such as the solvation free energy, surface area, volume, and so on, were calculated based on these most stable structures.

Step 5: Steps 1–4 were applied to the learning dataset and the test dataset, and then the neural network model was constructed from the learning set.

Step 6: The LogS values of the molecules in the test dataset were predicted by the MLR, ANN, and PLS models.

Preparation of data

The LogS data were selected from the previous reports.[23,24,33] Palmer et al.[23] included 1318 molecules, and 1290 molecules were used in Schwaighofer et al.[24] Most of the data originated from Fukunishi et al.[34] The 1290 molecules of reference 24 were included in the set of reference 23. Several of the 1318 molecules of reference 23 were counted twice. Among these molecules, some molecular structures could not be analyzed with our program, as they lacked descriptors. Finally, the total number of unique molecules used in this study was 1298. These molecular structures and their LogS values are summarized in the supporting information file. Most calculations for data preparation were performed using the myPresto program.[34] The two ion forms were generated by the Hgene/myPresto program. The atomic charges were calculated by the Gasteiger method in the Hgene program. A general Amber force field (GAFF) was used.[35]

Data on aggregators and non-aggregators were selected from the previous reports.[6,7] There were 56 and 57 aggregators and non-aggregators, respectively. In excess of 1000 aggregators are provided by the Shoichet group (http://shoichetlab.compbio.ucsf.edu/take-away.php). Since the ANN with WL method is time-consuming, we used the small set of aggregators. The 3D structures and the descriptors of these compounds were prepared in the same manner as the LogS data described above.

Results

Prediction accuracy without the weighted learning method

The efficiency of the descriptor set was evaluated using the MLR, ANN, and PLS models. The jackknife test was applied: all of the compounds were divided into two sets, a learning set with solubility data for machine learning and a test data set, whose LogS values the software should predict. The number of compounds in the learning set was 1198 (=1298−100), and the number in the test data set was 100. The molecules in the test data set were randomly selected. Ten pairs of these compound sets were prepared. Thus, a total of 1000 (=100 compounds × 10 trials) solubility data predictions were made.

The number of conformers is a parameter of our prediction method. We examined the conformer dependence of the predicted LogS value by using the ANN method. We examined the cases with a single conformer, 5 conformers, and 10

conformers. These conformers were randomly generated. The larger the number of conformers, the better the prediction result. The results with 5 conformers were similar to those with 10, and thus the number of conformers was set to 5 in the following study.

To compare our methods with the other prediction methods, we applied the MOE LogS predictor[36] and the Pipeline Pilot LogS predictor[37] to the same solubility data. The results are summarized in Table 2. The results obtained by the MLR, ANN, and PLS models were similar to each other, and the prediction results by our methods were similar to those obtained by the other predictors. These results suggest that our descriptors and the prediction methods were reasonably constructed.

Prediction accuracy with the weighted learning method

The WL method was applied to the MLR, ANN, and PLS models. By using these methods, the LogS values of a small number of molecules were estimated with and without the WL method. Two test data sets were used: the dataset used in the previous section and the dataset used in the "solubility challenge".[11,12] In both cases, the WL method worked well to improve the accuracy. The M and L_{max} values were set to 14 and 6, respectively.

The first test involved 33 molecules randomly selected from the learning set of 1298 compounds used in the previous section. For each test molecule, the learning set did not include the test molecule itself (one-leave-out test), hence the learning set consisted of 1297 (= 1298 – 1) molecules. Prediction with the WL method worked well. The R^2 value and the average error of the predicted LogS values of these 33 test molecules are summarized in Table 3. With all three models (MLR, ANN, and PLS), the WL method improved accuracy; the R^2 values were increased while the average and maximum errors were decreased. The R^2 value for the teaching set was not calculated by the ANN, since the ANN was too time consuming.

Table 2 Prediction accuracies for the teaching set and the test set

Method	Teaching set	Test set		
	R^2 value	R^2 value	Average error	Maximum error
MLR	0.864	0.846	0.81	2.62
ANN	0.949	0.906	0.63	3.36
PLS	0.856	0.846	0.82	2.70
Moe		0.884	2.26	5.54
Pipeline Pilot		0.847	2.12	4.97

Abbreviations: MLR, multi-linear regression; ANN, artificial neural network; PLS, partial least squares regression.

In the second test, the LogS values of the "solubility challenge" were predicted.[11,12] In the solubility challenge, the LogS values of 28 drug-like compounds must be predicted. Out of 32 compounds, 4 were too soluble to measure, so the other 28 LogS values must be predicted. The R^2 value, the average error, and the maximum error of LogS are summarized in Table 4. Again, the prediction results were better than the results without the WL method. Our prediction results summarized in Table 4 were worse than the results summarized in Table 3 but nevertheless were not as bad when compared to the results reported in reference 6. In reference 6, the range of R^2 values was 0.018 to 0.65 for these 28 compounds. Our R^2 values, around 0.5–0.6, were better. The percentage of entries that gave R^2 values better than 0.615 (obtained by the PLS model with the WL method) was 2% (2 entries out of 99). These results showed that our prediction methods were acceptable and could be used in the further study.

LogS values of aggregators and non-aggregators

Using all 1298 compounds with experimental LogS values as the learning set, we calculated the LogS values of the aggregators and non-aggregators. The prediction was performed using the MLR, ANN, and PLS models with the WL method. Figures 2a–c show the distribution of the LogS values of the aggregators and non-aggregators obtained by the MLR, ANN, and PLS models. The probability of aggregation at a LogS value was calculated by using Bayesian analysis.[38] The results were fitted by a sigmoid curve. Namely:

$$P_{agg}(LogS) = \frac{1}{1+e^{a(LogS-b)}} \quad (4)$$

where P_{agg}(LogS), a, and b are the probability of aggregation at the LogS solubility and two constants, respectively. The b value represents the LogS value at which the probability of aggregation is 50%. The a and b values obtained by the MLR were 1.477 and –4.911, respectively. The fitting error was 0.233. The a and b values obtained by the ANN were 1.013 and –4.274, respectively. The fitting error was 0.206. The a and b values obtained by the PLS were 1.354 and –4.784, respectively. The fitting error was 0.217.

Figure 3 shows the probability of aggregation vs the predicted LogS value. The aggregators and non-aggregators were clearly distinguished by the LogS value. In Figure 3, the percentage of aggregators reached 50% around the LogS value of –5. The MLR, ANN, and PLS models showed that the compounds with $-5 < \text{LogS} < -2$ are desirable as the non-aggregators.

Table 3 Prediction accuracies for 33 randomly selected molecules

Method		Teaching set	Test set	Average error	Maximum error
		R^2 value	R^2 value		
MLR	without WL	0.864	0.785	0.97	2.27
MLR	with WL	0.864	0.835	0.84	1.95
ANN	without WL		0.871	0.51	2.48
ANN	with WL		0.944	0.31	1.79
PLS	without WL	0.842	0.708	1.16	2.52
PLS	with WL	0.853	0.821	0.89	2.10

Abbreviations: MLR, multi-linear regression; ANN, artificial neural network; PLS, partial least squares regression.

The LogS distribution of the aggregators overlapped with that of the non-aggregators in the range of LogS < -5. Generally speaking, an aggregator has a small number of rotatable bonds, a flat, ring-like structure rich in nitrogen atoms, as suggested by a previous report.[8] We compared the chemical structures of the aggregators to those of the non-aggregators in this range, but found no clear difference.

We examined the frequency of aggregators in a compound library. The hetero atom dependence of aggregation was examined in two ways. One is the percentage of aggregation depending on the numbers of nitrogen and oxygen atoms of molecules. The other is the percentage of aggregation depending on the ratios of nitrogen and oxygen atoms of molecules. Nine subsets of the compound library were prepared in each. Each subset consisted of 55 compounds randomly selected from the library. The first group consisted of compounds with 200 Da $<$ MW $\leq$ 300 Da, the second group consisted of compounds with 300 Da $<$ MW $\leq$400 Da, and the third group consisted of compounds with 400 Da $<$ MW $\leq$500 Da. The first way in which we examined is the heteroatom-number dependence of aggregation. Each group was divided into three subsets. The first set consisted of a compound whose number of N/O atoms $<$5. The second set consisted of a compound whose number of N/O atoms $\geq$5 and $<$10. And the third set consisted of compound with number of N/O atoms $\geq$10. The second way in which we examined is the hetero-atom-ratio dependence of aggregation. Each group was divided into three subsets. The first set consisted of a compound whose ratio of N/O atoms $<$15%. The second set consisted of a compound with a ratio of N/O atoms $\geq$15% and $<$30%. And the third set consisted of a compound with a ratio of N/O atoms $\geq$30%. These compounds were randomly extracted from the LigandBox database.[39] The LogS values of these compounds were calculated by the PLS model with the WL method. The aggregator probability was calculated by equation (4).

The results are summarized in Tables 5 and 6. The probability of aggregation strongly depended on the molecular weight and the number and ratio of N/O atoms. The larger the compound is, the higher the probability of aggregation. Compounds with fewer N/O atoms were likely to be aggregates. The rule of five determines drug-likeness as the compound with the number of hydrogen bond acceptors (N/O atoms) must be no greater than 10. The results in Table 6 show the same trend as those in Table 5. Consideration of only this one rule of drug-like-

Table 4 Prediction accuracies for 28 molecules of the "solubility challenge"

Method		Teaching set	Test set	Average error	Maximum error
		R^2 value	R^2 value		
MLR	without WL	0.864	0.498	1.03	2.60
MLR	with WL	0.867	0.525	1.00	2.48
ANN	without WL		0.452	1.06	2.58
ANN	with WL		0.506	0.99	2.33
PLS	without WL	0.856	0.558	0.96	2.92
PLS	with WL	0.856	0.615	0.89	2.47
Pipeline Pilot	without WL		0.304	1.01	4.31
Moe	without WL		0.394	0.87	3.38
PLS[a]	without WL	0.712	0.497	1.36	3.37
PLS[a]	with WL	0.777	0.542	1.11	3.07

Note: [a]the LogS values of only 92 molecules provided by ref 37 were used as a teaching set.
Abbreviations: MLR, multi-linear regression; ANN, artificial neural network; PLS, partial least squares regression.

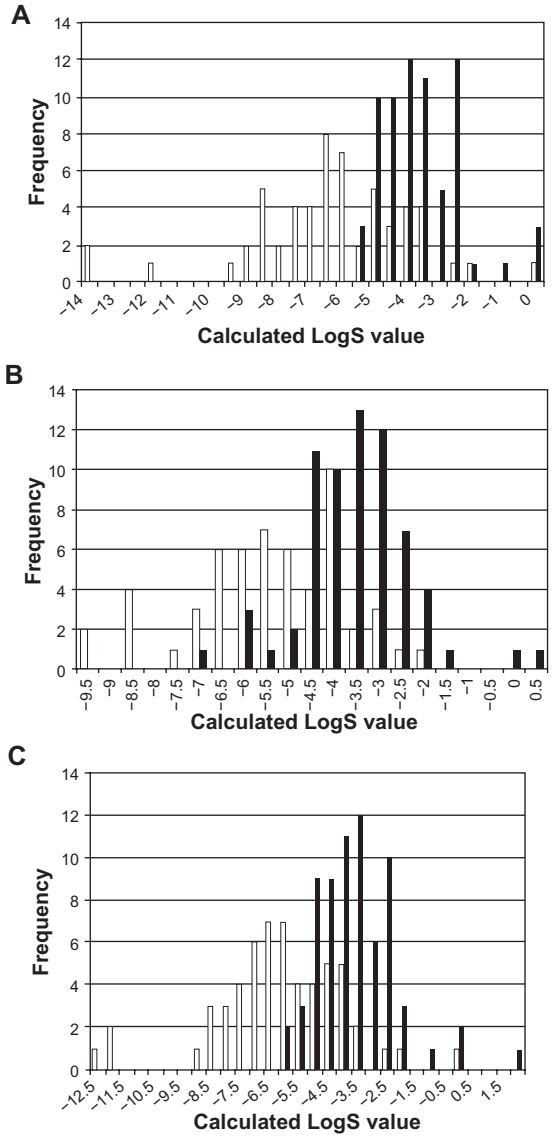

Figure 2 LogS values of aggregators and non-aggregators. Distribution of LogS values of aggregators and non-aggregators. White bars and black bars represent frequencies of aggregators and non-aggregators, respectively. **A**) MLR with the WL method. **B**) ANN with the WL method. **C**) PLS with the WL method.
Abbreviations: MLR, multi-linear regression; ANN, artificial neural network; PLS, partial least squares regression.

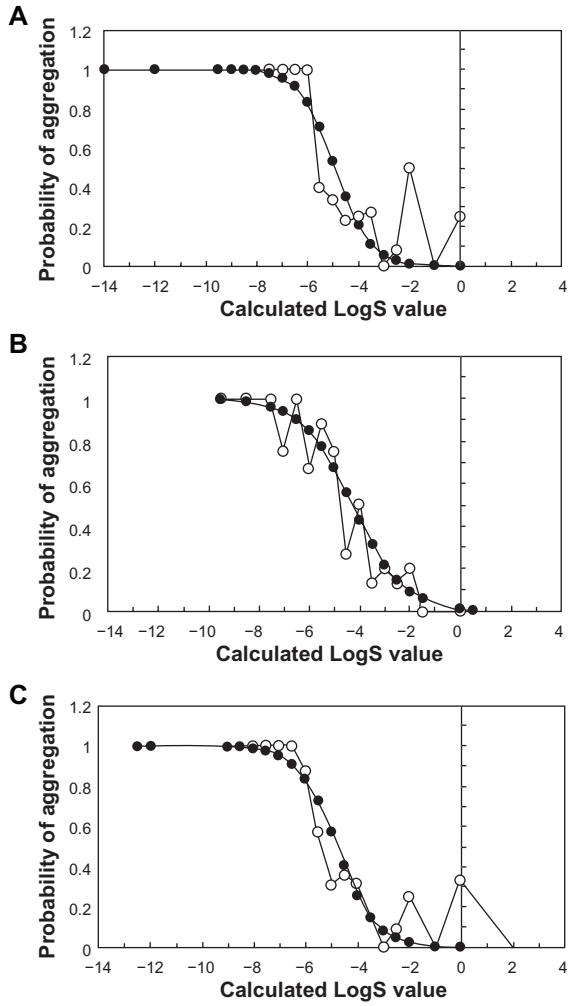

Figure 3 Probability of aggregation vs LogS value. **A**) MLR with the WL method. **B**) ANN with the WL method. **C**) PLS with the WL method.
Abbreviations: MLR, multi-linear regression; ANN, artificial neural network; PLS, partial least squares regression.

ness increases the chance of aggregation. The other condition, which is that the LogP value must be <5, increases the LogS value by equation (1) and reduces the chance of aggregation.

The reliability of the values in Tables 5 and 6 is unclear. In our teaching set, only 157 out of 1298 compounds (12.1%) show LogS < −5; the other 87.9% of the compounds show LogS > −5. The probability of aggregation was estimated to be high especially when the LogS value is <−5. The number of low solubility compounds was small. There were data on only 56 aggregators, which is still small.

Discussion

In the present study's definition of aggregators, the molecule inhibits several target proteins. The actual reasons for aggregation include micelle colloid formation, nonspecific interactions with proteins, and other, unknown mechanisms. Some aggregators contain N-N bonds, which are chemically unstable, and thus chemical reactivity could be another cause of aggregation. Such compounds should be filtered out by software. A molecule with a low LogS value should

Table 5 Percentage of aggregators of compound set depending on number of nitrogen/oxygen (N/O) atoms

		MW (Da)		
		200–300	300–400	400–500
No of N/O	0–4	23.0	62.6	92.0
	5–9	9.9	45.1	69.8
	10–	3.2	18.9	36.6

Table 6 Percentage of aggregators of compound set depending on ratio of nitrogen/oxygen atoms

		MW (Da)		
		200–300	300–400	400–500
N/O%	0–15%	38.9	72.6	91.1
	15–30%	13.3	42.8	66.2
	30–100%	7.0	28.8	39.1

easily form a micelle colloid, but low solubility is not the only reason for nonspecific protein-compound interactions. If we applied multiple-target screening (MTS)[40] as an *in silico* screening method, it could remove the nonselective compounds. In the MTS method, each compound is docked to many proteins, including a target protein. Then, the compounds that show the strongest affinities with the target protein, among the other different proteins, are selected as the hit compounds. If the LogS prediction is applied to the hit compounds by the MTS method, then the number of aggregators could be reduced.

In previous reports aggregators were predicted using the substructures of compounds.[1,2] These predictions worked well, but the physical meaning of the substructures was unclear. The clogP value was used in the prediction, and the result suggested that compounds with high clogP values (hydrophobic) are likely to aggregate.[1,2] This insight is consistent with our result that the aggregators show low LogS values and with the experimental finding that the aggregators form micelle colloids in water.[3]

Conclusion

We introduced the weighted learning (WL) method in the prediction of LogS by the MLR, ANN, and PLS models. With the WL method, the LogS predictor studies the learning set by focusing on the molecules that are similar to the query molecule. The WL method can achieve high prediction accuracy and can be combined with the MLR, ANN, and PLS models.

We applied our method to the prediction of aggregators and non-aggregators. The non-aggregators showed higher LogS values than the aggregators. One of the useful thresholds is LogS = −5. The probability of aggregation was given by a simple sigmoid function of LogS (see equation 4). The molecules with LogS $<$ −5 were potential aggregators, while those with LogS $>$ −5 were potential non-aggregators. One of the reasons for aggregation is low solubility. In the lead optimization process, we recommend that the lead compounds satisfy the condition of LogS $>$ −5. We also showed a simple look-up table to estimate the percent of aggregation depending on the molecular weight and the ratio of nitrogen/oxygen atoms. The percentage of aggregation strongly depended on the molecular weight, the number and ratio of nitrogen/oxygen atoms. This knowledge will help in the design of a library for drug screening.

Acknowledgments

This work was supported by the New Energy and Industrial Technology Development Organization of Japan (NEDO) and by the Ministry of Economy, Trade, and Industry (METI) of Japan. Dr. Sumie Tajima (Beyond Computing Co. Ltd.) supported the development of the software.

Supporting information available

This article contains supplementary material (lists of LogS values and the compounds provided in the SMILES format) that is available at http://presto.protein.osaka-u.ac.jp/myPresto4/ and http://medals.jp/myPresto/index.html.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Jadhav A, Ferreira RS, Klumpp C, et al. Quantitative analysis of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. *J Med Chem*. 2010;53:37–51.
2. Shoichet BK. Screening in a spirit haunted world. *Drug Discov Today*. 2006;11:607–615.
3. Feng BY, Simeonov A, Jadhav A, et al. A high-throughput screen for aggregation-based inhibition in a large compound library. *J Med Chem*. 2007;50:2385–2390.
4. Ryan AJ, Gray NM, Lowe PN, et al. Effect of detergent on "promiscuous" inhibitors. *J Med Chem*. 2003;46:3448–3451.
5. McGovern SL, Helfand BT, Feng B, et al. A specific mechanism of nonspecific inhibition. *J Med Chem*. 2003;46:4265–4272.
6. Roche O, Schneider P, Zuegge J, et al. Development of a virtual screening method for identification of "frequent hitters" in compound libraries. *J Med Chem.* 2002;45:137–142.
7. Crisman TJ, Parker CN, Jenkins JL, et al. Understanding false positives in reporter gene assays: in silico chemogenomics approaches to prioritize cell-based HTS data. *J Med Chem*. 2007;47:1319–1327.
8. Seidler J, McGovern SL, Doman TN, et al. Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J Med Chem*. 2003;46:4477–4486.
9. Rao H, Li Z, Li X, et al. Identification of small molecule aggregators from large compound libraries by support vector machines. *J Comput Chem*. 2010;31:752–763.
10. Hughes LD, Palmer DS, Nigsch F, et al. Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and LogP. *J Chem Inf Comput Sci*. 2008;48:220–232.
11. Llinas A, Glen RC, Goodman JM. Solubility challenge: Can you predict solubilities of 32 molecules using a database of 100 reliable measurements? *J Chem Inf Comput Sci*. 2008;48:1289–1303.
12. Hopfinger AJ, Esposito EX, Llinas A, et al. Findings of the challenge to predict aqueous solubility. *J Chem Inf Comput Sci*. 2009;48:1–5.

13. Wakita K, Yoshimoto M, Miyamoto S, et al. A method for calculation of the aqueous solubility of organic compounds by using new fragmental solubility constants. *Chem Pharm Bull*. 1986;34:4663–4681.
14. Suzuki T. Development of an automatic estimation system for both the partition coefficient and aqueous solubility. *J Comput Aided Mol Des*. 1991;5:149–166.
15. Kuhne R, Ebert RU, Kleint F, et al. Group contribution methods to estimate water solubility of organic chemicals. *Chemosphere*. 1995;30:2061–2077.
16. Jorgensen WL, Duffy EM. Prediction of drug solubility from Monte Carlo simulations. *Bioorg Medicinal Chem Lett*. 2000;10:1155–1158.
17. Ran Y, Yalkowsky SH. Prediction of drug solubility by the general solubility equation (GSE). *J Chem Inf Comput Sci*. 2001;41:354–357.
18. Yan A, Gasteiger J. Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *J Chem Inf Comput Sci*. 2003;43:429–434.
19. Sum H. A universal molecular descriptor system for prediction of LogP, LogS, LogBB, and absorption. *J Chem Inf Comput Sci*. 2004;44:748–757.
20. Bergstrom CAS, Wassvik CM, Norinder U, et al. Global and local computational models for aqueous solubility prediction of drug-like molecules. *J Chem Inf Comput Sci*. 2004;44:1477–1488.
21. Huuskonen J, Salo M, Taskinen J. Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *J Chem Inf Comput Sci*. 1998;38:450–456.
22. Wang J, Krudy G, Hou T, et al. Development of reliable aqueous solubility models and their application in druglike analysis. *J Chem Inf Comput Sci*. 2007;47:1395–1404.
23. Palmer DD, O'Boyle NM, Glen RC, et al. Random forest models to predict aqueous solubility. *J Chem Inf Comput Sci*. 2007;47:150–158.
24. Schwaighofer A, Schroeter T, Mika S, et al. Accurate solubility prediction with error bars for electrolytes: a machine learning approach. *J Chem Inf Comput Sci*. 2007;47:407–424.
25. Obrezanova O, Gola JM, Champness EJ, et al. Automatic QSAR modeling of ADME properties: blood-brain barrier penetration and aqueous solubility. *J Comput Aided Mol Des*. 2008;22:431–440.
26. Hilal SH, Saravanaraj AN, Whiteside T, et al. Calculating physical properties of organic compounds for environmental modeling from molecular structure. *J Comput Aided Mol Des*. 2007;21:693–708.
27. Yan A, Gasteiger J, Krug M, et al. Linear and nonlinear functions on modeling of aqueous solubility of organic compounds by two structure representation method. *J Comput Aided Mol Des*. 2004;18:75–87.
28. Zupan J, Gasteiger J. Neural networks in chemistry and drug design 2nd edition; Wiley-VCH; 1999.
29. Joback KG, Reid RC. Estimation of pure-component properties from group-contributions. *Chem Eng Comm.* 1987;57:233–243.
30. Tajima S et al. Computed property data base: CPDB – development of an integrated software tool for molecular design: MolWorks. *J Comput Chem Jpn.* 2006;5:23–28.
31. Still WC, Tempczyk A, Hawley RC, et al. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc.* 1990;112:6127–6129.
32. Hawkins DG, Cramer JC, Truhlar GD. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J Phys Chem.* 1996;100:19824–19839.
33. Huuskonen J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J Chem Inf Comput Sci.* 2000;40:773–777.
34. Fukunishi Y, Mikami Y, Nakamura H. The filling potential method: A method for estimating the free energy surface for protein-ligand docking. *J Phys Chem B.* 2003;107:13201–13210.
35. Wang J, Wolf RM, Caldwell JW, et al. Development and testing of a general amber force field. *J Compt Chem.* 2004;25:1157–1174.
36. Hou TJ, Xia K, Zhang W, et al. ADME Evaluation in drug discovery. 4 Prediction of aqueous solubility based on atom contribution approach. *J Chem Inf Comput Sci.* 2004;44:266–275.
37. Tetko I, Tanchuk YV, Kasheva T, Villa A. Estimation of aqueous solubility of chemical compounds using E-state indices. *J Chem Inf Comput Sci*. 2001;41:1488–1493.
38. Jacobsson M, Liden P, Stjernschantz E, Bostrom H, Norinder U. Improving structure-based virtual screening by multivariate analysis of scoring data. *J Med Chem*. 2003;46:5781–5789.
39. Fukunishi Y, Sugihara Y, Mikami Y, Sakai K, Kusudo H, Nakamura H. Advanced in-silico drug screening to achieve high hit ratio-development of 3D-compound database. *Synthesiology*. 2009;2:60–68.
40. Fukunishi Y, Mikami Y, Kubota S, Nakamura H. Multiple target screening method for robust and accurate in silico screening. *J Mol Graph Model*. 2005;25:61–70.