Are There Differences between Launched Drugs, Clinical Candidates, and Commercially Available Compounds?

Kazuki Ohno,[†,‡] Yuichi Nagahara,[‡,§] Kazuhisa Tsunoyama,[†,‡] and Masaya Orita*[,†,‡]

Drug Discovery Research, Astellas Pharma Inc., 21 Miyukigaoka, Tsukuba, Ibaraki 305-8585, Japan, and
School of Political Science and Economics, Meiji University, 1-1 Kanda-Surugadai, Chiyoda-ku,
Tokyo, 101-8301, Japan

Received January 14, 2010

To clarify the differences between commercially available compounds, clinical candidates, and launched drugs with regard to distribution of physicochemical properties and to characterize the correlation between physicochemical properties, we conducted analyses on physicochemical descriptors of commercially available compounds, clinical candidates, and launched drugs. Initial analysis of the marginal distribution of each physicochemical property showed that the distribution of commercially available compounds obeys a more normal distribution than that of launched drugs and clinical candidates. In addition, we calculated correlation coefficient values between values of physicochemical properties and found little similarity between values of clinical candidates and those of commercially available compounds, while observing marked similarity between values of clinical candidates and those of launched drugs. We also analyzed joint distribution for two physicochemical properties, with results showing that, similar to marginal distribution, the joint distribution of commercially available compounds obeys a more normal distribution than that of launched drugs and clinical candidates. We then assessed items using the Nagahara method, originally developed by one of this study's authors. Results showed that the probability distribution of molecular weight and log *P* for commercially available compounds was much narrower than that of launched drugs and clinical candidates. In conclusion, clinical candidates are more similar to launched drugs than to commercially available compounds with regard to marginal distribution, joint distribution, and correlation coefficients. These findings provide deeper insight regarding the concept of "druglikeness".

INTRODUCTION

Compound collections are one of the most important research assets for a pharmaceutical company, as high-throughput screening (HTS) of a compound collection provides starting points for drug discovery. Designing such a compound library is therefore an extremely important task for computational chemists. When adding novel compounds to an in-house chemical collection, several points must be taken into consideration: compound novelty, chemical space coverage, structural diversity, and desirable physicochemical properties. Many computational chemists have therefore spent considerable time and effort developing novel concepts and methods for this precise purpose.

In 1997, the Lipinski group, working for Pfizer, introduced the concept of "druglikeness". They proposed the use of a physicochemical property-based filter, dubbed the "rule of five", which sets the upper limits of descriptors as follows: molecular weight (MW) = 500 Da, calculated logarithm of the octanol/water partition coefficient (log *P*) = 5, number of hydrogen-bond donors (HBD) = 5, and number of hydrogen-bond acceptors (HBA) = 10.[1] Since this method's publication, the usefulness of physicochemical properties in constructing chemical libraries has been investigated to further the understanding of the characteristics of chemical libraries.

Veber et al. proposed the additional criteria of "number of rotatable bonds (RTB) ≤ 10" and "polar surface area (PSA) < 140 Å^2".[2] To aid fragment library design in the context of fragment-based drug discovery, researchers at Astex recently introduced the "rule of three".[3] In addition, our group proposed two novel indices based on the "golden ratio".[4] Further, a research group lead by Oprea has published several excellent studies on the difference between drugs, lead compounds, and other compounds with regard to physicochemical properties.[5−7]

Interestingly, several distributions of physicochemical properties reported by the Oprea group obey asymmetry and fat-tailed (non-normal) distribution. Analysis of such non-normal distributions has recently drawn attention as an important research topic, and the Pearson distribution system has found use in a range of fields, including financial engineering, econometrics, education, psychology, natural science, and technology. For example, in the field of financial engineering, non-normal distributions, particularly those of the Pearson distribution system, are applied to stock returns and stochastic volatility models.[8−13] In signal processing, one of the authors (Y.N.) examined the blind separation problem using independent component analysis.[14] In the field

* Corresponding author phone: +81-29-863-6756; fax: +81-29-852-5391; e-mail: masaya.orita@jp.astellas.com.
[†] Astellas Pharma Inc.
[‡] Authors contributed equally to this work.
[§] Meiji University.

10.1021/ci100023s
Published on Web 04/15/2010

of pharmaceuticals, the Pearson distribution system has been used to assess the issue of conformational sampling in proteins.[15]

Here, we investigated the distribution of physicochemical properties of chemical libraries. Furthermore, we calculated correlation coefficient values between chemical libraries. Finally, we have proposed the novel analysis method based on the shape of descriptor and correlation.

THEORY

The normal distribution plays a central role in statistics, yet empirical distributions of data do not always follow a normal distribution. To describe these distributions, therefore, several non-normal distributions have been derived and introduced. However, given the variety of parametric distributions, no single parametric distribution can be expected to represent the wide range of kurtosis and skewness possible. Researchers have therefore developed distribution systems considering several specific distributions, the most famous of which is the Pearson distribution system.

Pearson[16] defined the Pearson distribution system using the following differential equation with respect to the probability density function, p:

$$-\frac{\frac{\mathrm{d}p}{\mathrm{d}x}}{p(x)} = \frac{b_0 + b_1 x}{c_0 + c_1 x + c_2 x^2} \tag{1}$$

Many statistics researchers have referred to this distribution system in their own studies.[17−19] One of the present authors (Y.N.) overcame certain difficulties inherent in using the Pearson type IV distribution, both skewed and fat-tailed, adapting it for practical use, and estimated parameters for stock returns using the maximum likelihood method.[10] In more recent years, a practical approach to using nearly all types of Pearson distribution systems, including the type IV distribution for non-Gaussian filters, was developed.[11] These previous studies have facilitated the practical application of the Pearson distribution system.

However, few practical methods of generating multivariate non-normal distributions free from analytical restrictions using the parametric distribution system have been developed thus far. One of the authors (Y.N.) used both the method of moments and the maximum likelihood method to develop a method based on random numbers for constructing multivariate non-normal distributions free from analytical restrictions.[20]

In the present study, the maximum likelihood method is used for estimating the parameters of multivariate non-normal distributions whose density functions are numerically approximated by random numbers. Given that the present method requires examination of a large number of parameter grids, focusing on the skewness and kurtosis of the Pearson distribution system in order to maximize the likelihood function, this method is extremely computational intensive. However, computing speeds continue to increase, as do the number of CPU multicores, thereby enabling quicker random number generation using methods such as the Mersenne Twister.[21] Such advances have aided greatly in increasing the practicality of applying our method.

A brief explanation about the theory is the following. (1) We generate random numbers from two standardized independent random variables, ξ_1 and ξ_2, which belong to the Pearson distribution system and have certain skewness and certain kurtosis. "Standardized" means that its average and its standard deviation equal to zero and one, respectively. (2) Using $\mathbf{T}$ ($= \mathbf{\Sigma}^{1/2}$) calculated by the covariance matrix of the standardized sample data and random numbers above, we obtain random vector $\mathbf{X}$ ($= \mathbf{T}\xi$) represented by random numbers. The joint probability distribution of random vector $\mathbf{X}$ is the multivariate non-normal distribution or the multivariate normal distribution. (3) We divide the coordinate planes by a step. By dividing a number of points of random numbers in each division by a number of total points of random numbers, we obtain the probability and we calculate the log-likelihood of the standardized sample data. (4) Changing the combination of the skewness and the kurtosis, we repeat steps 1−3. By using the grid search, we obtain the estimated skewness and kurtosis at the maximum log-likelihood. Finally, by the maximum likelihood method, we are able to obtain the multivariate distribution fitted to the sample data.

METHODS

Data Set Preparation. Three chemical libraries were prepared.[1] For the commercially available compound library (CAC), 137 013 compounds were randomly selected from the ZINC purchasable database (total available, 13 669 837 compounds), which is a commercially available compound database provided by the Shoichet Laboratory in the Department of Pharmaceutical Chemistry at the University of California (San Francisco, CA).[22] For the clinical candidate library (CC), 8955 clinical candidates were obtained from the clinical candidate database provided by GVK BioSciences (Hyderabad, India).[23] For the launched drug library (LD), 3767 compounds were obtained from the drug database provided by GVK BioSciences.

Descriptor Selection and Computation of Chosen Descriptors. Using multiple descriptors to describe a single compound may lead to system overdescription. We therefore attempted to use diverse and at least partially independent descriptors. Furthermore, we are very interested in druglike features. The factors in Lipinski's rule of five seem to be a good starting point. Thus, our descriptor sets are MW, log P, HBA, HBD, RTB, and PSA. We used Pipeline Pilot version 7.5.2 provided by Accelrys Software Inc. to calculate the six chosen descriptors. In this program, log P was calculated using the AlogP calculation method developed by Ghose et al.[24]

The Pearson Distribution System and Pearson Type VII and IV Distributions. Pearson distribution system types are defined by the κ value, which is defined as

$$\kappa = \frac{\beta_1(\beta_2 + 3)^2}{4(2\beta_2 - 3\beta_1 - 6)(4\beta_2 - 3\beta_1)} \tag{2}$$

where β_1 and β_2 represent squared skewness and kurtosis, respectively.

For $\kappa < 0$, $0 < \kappa < 1$, and $\kappa > 1$, the distribution is called type I, IV, and VI, respectively, and are known as the main types.[17] Types VII and II, and normal distribution are included for $\kappa = 0$.

In this paper, the probability density function of Pearson type IV is defined by

$$p(x) = \frac{\Gamma(b+b\delta \mathrm{i})\Gamma(b-b\delta \mathrm{i})\tau^{2b-1}}{\Gamma(b)\Gamma\left(b-\frac{1}{2}\right)\pi^{1/2}} \frac{\exp\left\{2b\delta \arctan\left(\frac{x-\mu}{\tau}\right)\right\}}{\{(x-\mu)^2+\tau^2\}^b} = \frac{\Gamma(b+b\delta \mathrm{i})\Gamma(b-b\delta \mathrm{i})}{\{\Gamma(b)\}^2} \times \frac{\tau^{2b-1}}{\mathrm{B}\left(b-\frac{1}{2},\frac{1}{2}\right)} \times \frac{\exp\left\{2b\delta \arctan\left(\frac{x-\mu}{\tau}\right)\right\}}{\{(x-\mu)^2+\tau^2\}^b} \qquad (3)$$

where μ, τ, b, and δ represent the location, scale, shape, and noncentral parameter and Γ, B, and i indicate gamma function, beta function, and imaginary number, respectively. The moments are as follows:

$$\text{mean} = \frac{b\delta\tau}{b-1} + \mu \qquad (4)$$

$$\text{variance} = \frac{\tau^2}{2b-3}\left\{1 + \left(\frac{b}{b-1}\delta\right)^2\right\} \qquad (5)$$

$$\text{skewness} = \frac{2b(2b-3)^{1/2}}{(b-1)(b-2)} \frac{\delta}{\left\{1 + \left(\frac{b}{b-1}\delta\right)^2\right\}^{1/2}} \qquad (6)$$

$$\text{kurtosis} = \frac{3(2b-3)}{(2b-5)} \frac{\left\{1 + \frac{b+2}{b-2}\left(\frac{b\delta}{b-1}\right)^2\right\}}{\left\{1 + \left(\frac{b}{b-1}\delta\right)^2\right\}} \qquad (7)$$

When $\delta = 0$, the distribution is characterized as Pearson type VII. In the Pearson distribution system, provided the type is specified, its parameters of distribution can be determined on the basis of the mean, variance, skewness, and kurtosis. The typical relationship between squared skewness and kurtosis has been described in textbooks.[17−19] More practical methods of implementing the Pearson distribution system and generating random numbers using this system, including the type IV distribution, have been introduced by Nagahara.[20]

Multivariate Non-Normal Distributions (Nagahara Method). According to Nagahara,[12,20] multivariate non-normal distributions are constructed with $\xi_1, ..., \xi_m$ as independent random variables and $E(\xi_j) = 0$, $E(\xi_j^2) = 1$, $E(\xi_j^3) = \zeta_j$, $E(\xi_j^4) = \kappa_j$, and $\boldsymbol{\xi} = (\xi_1, ..., \xi_m)'$. $\mathbf{T} = (t_{ij})$ is a nonrandom $p \times m$ matrix of rank p such that $\mathbf{TT}' = \mathbf{\Sigma}$ and $m \geq p$. We note that $\mathbf{T} = \mathbf{\Sigma}^{1/2}$ when $m = p$, giving the random vector

$$\mathbf{X} = \mathbf{T}\boldsymbol{\xi} \qquad (8)$$

which generally follows a nonelliptical distribution with $\mathrm{Cov}(\mathbf{X}) = \mathbf{\Sigma}$. The respective marginal skewness and kurtosis of x_i are derived by

$$\mathrm{skew}(x_i) = \sum_{j=1}^{m} \frac{t_{ij}^{\ 3}\varsigma_j}{\sigma_{ii}^{\ 3/2}} \qquad (9)$$

and

$$\mathrm{kurt}(x_i) = \left\{\sum_{j=1}^{m} \frac{t_{ij}^{\ 4}(k_j-3)}{\sigma_{ii}^{\ 2}} + 3\right\} \qquad (10)$$

where t_{ij} are the elements of $\mathbf{T}$ and σ_{ii} are the elements of $\mathbf{\Sigma}$. According to Nagahara,[20] the maximum likelihood method is better able to estimate the parameters than the method of moments. We therefore propose implementing the maximum likelihood method by approximating the density function based on random numbers generated by our method. Using these random numbers, we may then numerically approximate the density function. The details of this method have been previously described by Nagahara.[20]

Table 1. Descriptive Statistics of Six Physicochemical Descriptors

	database	number	average	SD	skewness	kurtosis
MW	CAC	137013	391.05	75.23	–0.61	3.10
	CC	8955	421.89	151.22	0.99	4.35
	LD	3767	333.87	141.65	1.37	6.31
log *P*	CAC	137013	3.32	1.48	–0.19	3.86
	CC	8955	2.99	2.52	–0.38	5.33
	LD	3767	2.31	2.51	–0.68	7.07
HBA	CAC	137013	4.53	1.60	0.30	3.41
	CC	8955	5.50	2.97	1.64	7.80
	LD	3767	4.57	3.18	2.25	12.19
HBD	CAC	137013	1.21	0.89	0.78	4.87
	CC	8955	2.14	1.95	2.20	12.09
	LD	3767	1.74	1.94	2.77	15.48
RTB	CAC	137013	6.06	2.35	0.20	3.11
	CC	8955	6.96	5.09	1.86	9.48
	LD	3767	5.23	4.11	2.09	12.67
PSA	CAC	137013	89.87	32.90	0.25	3.13
	CC	8955	99.64	59.00	1.73	8.11
	LD	3767	79.60	60.00	2.15	11.12

[a] MW, molecular weight; HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; RTB, number of rotatable bonds; PSA, polar surface area; CAC, commercially available compounds; CC, clinical candidates; LD, launched drugs; SD, standard deviation.

Correlation Coefficient Values between Two Descriptors. The correlation coefficient values between two descriptors were calculated using Pipeline Pilot (Accelrys Inc., San Diego, CA). Because we considered six descriptors (MW, log *P*, HBA, HBD, RTB, and PSA) in the present study, a total of 15 correlation coefficient values (${}_6\mathrm{C}_2 = 15$; “correlation matrix”) were obtained for each data set.

The correlation matrix includes data regarding the physicochemical characteristics of compounds in a chemical library. If the correlation matrix of library A is similar to that of library B, compounds in library A are likely similar to those in library B. Therefore, a library with a correlation matrix similar to those of marketed drugs presumably includes many druglike compounds. To evaluate whether or not the correlation matrix of a library was similar to that of LD, we introduced the druglike score of the library (DSL), defined as follows

$$\mathrm{DSL} = \frac{\left(\sum_{i \neq j} |C_{i,j,\mathrm{target}} - C_{i,j,\mathrm{LD}}|\right)}{2 \times 15} \qquad (11)$$

where $C_{i,j,\mathrm{target}}$ and $C_{i,j,\mathrm{LD}}$ represent the correlation coefficient values between descriptors i and j in the target chemical library and drug database, respectively.

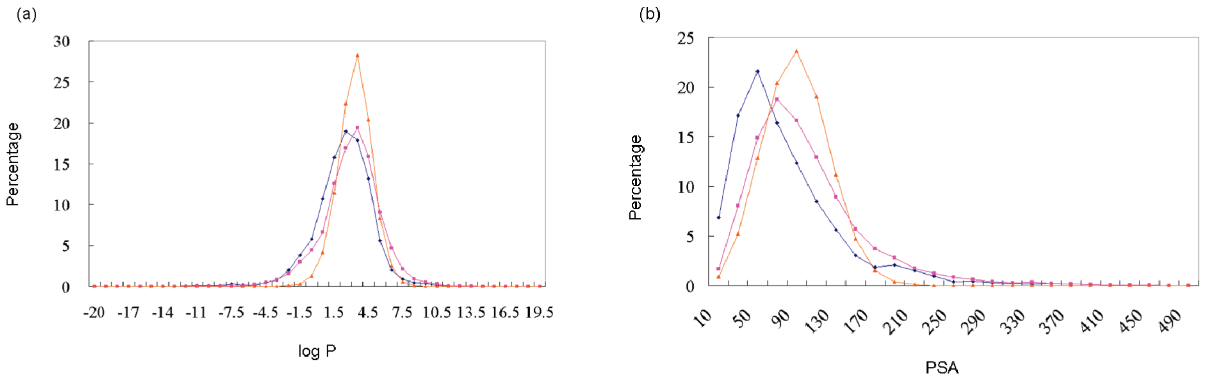


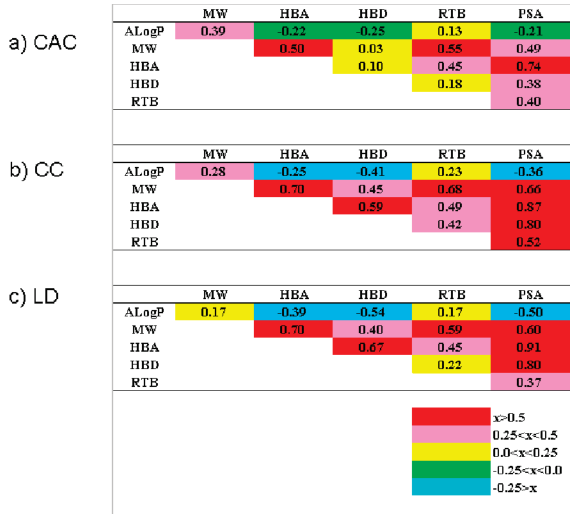
Figure 1. Probability distribution of log *P* (a) and MW (b).

RESULTS AND DISCUSSION

Marginal Distribution. The skewness and kurtosis values for normal distribution are 0 and 3, respectively. Skewness, kurtosis, average, and standard deviation values for the six descriptors across three chemical libraries are described in Table 1. With regard to each descriptor, the respective skewness and the kurtosis values in CAC were closer to 0 and 3 than those in the CC or drug libraries, indicating that the distribution of commercially available compounds more closely obeys a normal distribution than that of launched drugs or clinical candidates. Physicochemical descriptors in commercially available compounds may be expected to demonstrate more randomness than those in launched drugs or clinical candidates because commercially available compounds are derived using multiple constituent compounds from different suppliers. In contrast, launched drugs and clinical candidates must satisfy requirements such as high solubility and high stability.

The observed non-normality in physicochemical property distributions may arise due to the druglikeness features of launched drugs and clinical candidates. Distributions of log *P* and PSA are shown in Figure 1a,b. The distributions of both parameters in CAC obey a normal distribution (orange lines), while log *P* distribution for LD and CC data sets are asymmetrical and fat-tailed to the left (blue and pink lines, respectively), indicating the existence of a large number of compounds with extremely low log *P*. In detail, the ratios of compounds with extremely low log *P* (log $P < -2$) were 3.4% (302 out of 8955), 4.4% (166 out of 3767), and 0.17% (235 out of 137013) for LD, CC, and CAC libraries, respectively. From these data, we can see that the percentage of LD and CC with extremely low log *P* values was over 20 times that of CAC. Further, PSA values for LD and CC followed an asymmetrical distribution that was fat-tailed to the right (blue and pink lines, Figure 1b). The ratio of compounds with extremely high PSA (PSA > 200) were 6.1% (544 out of 8955), 4.6% (175 out of 3767), and 0.14% (196 out of 137013) for LD, CC, and CAC, respectively. From these data, we can see that the percentage of LD and CC with extremely high PSA values was over 30 times that of CAC. These findings demonstrate that the physicochemical properties of launched drugs and clinical candidates often have abnormal values, suggesting that adding compounds with abnormal physicochemical properties to the HTS library may be reasonable.

On starting the present study, we hypothesized that the average value of physicochemical features of clinical candidates more closely resembled those of launched compounds than those of commercially available compounds, due to their druglike properties. With regard to MW, HBA, RTB, and PSA, however, results showed that the average values of these parameters for LD more closely resembled those of CAC than they resembled CC (Table 1), shedding doubt on the hypothesis that average values of LD are similar to those of CC.


a) CAC

	MW	HBA	HBD	RTB	PSA
ALogP	0.39	-0.22	-0.25	0.13	-0.21
MW		0.50	0.03	0.55	0.49
HBA			0.10	0.45	0.74
HBD				0.18	0.38
RTB					0.40

b) CC

	MW	HBA	HBD	RTB	PSA
ALogP	0.28	-0.25	-0.41	0.23	-0.36
MW		0.70	0.45	0.68	0.66
HBA			0.59	0.49	0.87
HBD				0.42	0.80
RTB					0.52

c) LD

	MW	HBA	HBD	RTB	PSA
ALogP	0.17	-0.39	-0.54	0.17	-0.50
MW		0.70	0.40	0.59	0.60
HBA			0.67	0.45	0.91
HBD				0.22	0.80
RTB					0.37

🟥	x>0.5
🟪	0.25<x<0.5
🟨	0.0<x<0.25
🟩	-0.25<x<0.0
🟦	-0.25>x

Figure 2. Correlation matrix between six descriptors. Red, pink, yellow, green, and blue indicate correlation coefficients of $> = 0.5$, $0.25 = < c < 0.5$, $0 = < c < 0.25$, $-0.25 = < c < 0$, and < -0.25, respectively.

Of particular note is the fact that skewness and kurtosis values are similar between clinical candidates and launched drugs. For example, with MW, the skewness values for CC and LD are both positive (0.99 and 1.37), whereas that for CAC is negative (−0.61). Further, the kurtosis value for CAC is approximately 3, whereas those for CC and LD are both over 4. Such information regarding the shape of distribution (skewness and kurtosis values) provides researchers with deeper insight into druglikeness.

We also noted that the average value of each descriptor in LD is larger than that in CC. Specifically, the average values of MW, log *P*, HBA, HBD, RTB, and PSA in CC are larger than those in LD by 88, 0.32, 0.93, 0.40, 1.73, and 20.0, respectively, results which are consistent with findings from previous studies.[5−7]

Correlation Matrix. The correlation matrix of physicochemical descriptors is shown in Figure 2. Interestingly, the correlation matrix of LD is considerably similar to that of

CC while dramatically differing from that of CAC. Such a similarity may be due to the druglikeness features of launched drugs and clinical candidates.

In general, the MW of a compound correlates with the log *P*, because increasing MW often induces a subsequent increase in hydrophobicity. In fact, a significant correlation can be seen between MW and log *P* (0.39) in CAC. In contrast, this correlation is relatively weak in LD (0.17) and CC (0.28), an observation which may be explained as follows. In general, medicinal chemists attempt to synthesize compounds with high target affinity (leading to compounds with high MWs) and high solubility (leading to compounds with low log *P*) during optimization, thereby reducing or altogether destroying the above-mentioned correlation. As a result, in the case of clinical candidates and launched drugs, the correlation tends to be weak, although the correlation in commercially available compounds is relatively high.

Little correlation has been noted between HBA and HBD in CAC (0.10), whereas correlation between these parameters is significant in LD (0.67) and CC (0.59). Similarly, little correlation has been noted between MW and HBD in CAC (0.03), whereas correlation between these parameters is significant in LD (0.45) and CC (0.40). Such differences in correlation may also be associated with the druglikeness features of clinical candidates and launched drugs.

Of note is the fact that the ratio of compounds with high HBA (>10) and high HBD (>5) in CAC is significantly lower (39 compounds, 0.028%) than the ratios in LD (95 compounds, 1.1%) and CC (217 compounds, 0.69%), with ratios in LD and CC being 20−40 times that of CAC. With regard to LD and CC, compounds with high HBA tend to have high HBD values because of the strong correlation between HBA and HBD. Compounds with such high HBA and HBD values are therefore quite common among LD and CC. For CAC, however, compounds with high HBA tend to have relatively low HBD values because of the weak correlation between HBA and HBD, resulting in compounds with high HBA and high HBD being quite rare among CAC. Such correlations between physicochemical properties can provide researchers with a significant wealth of data.

Joint Distribution. Skewness and kurtosis values for ξ_1 and ξ_2 for joint distribution are shown in Table 2. The values for CAC more closely resemble the respective normal values of 0 and 3 than do values for CC and LD, indicating that the joint distribution of commercially available compounds more closely obeys a normal distribution than does that of launched drugs or clinical candidates. Interestingly, the skewness and kurtosis values for ξ_1 and ξ_2 among CC are almost identical to those in LD. Taking these findings into account, we now understand that clinical candidates more closely resemble launched drugs than they resemble commercially available compounds with regard to both marginal and joint distribution.

Table 2. Correlation Coefficient, ξ_1, and ξ_2 Values for Joint Distribution[a]

	database	correlation		skewness	kurtosis
MW:log *P*	CAC	0.39	ξ_1	–0.5	3
			ξ_2	0	3
	CC	0.28	ξ_1	0.75	5
			ξ_2	–0.75	7
	LD	0.17	ξ_1	0.75	
			ξ_2	–0.75	9
HBA:HBD	CAC	0.10	ξ_1	0.5	5
			ξ_2	0.75	3
	CC	0.59	ξ_1	0.75	
			ξ_2	0.75	7
	LD	0.67	ξ_1	0.75	
			ξ_2	0.75	7
RTB:PSA	CAC	0.40	ξ_1	0	3
			ξ_2	0	3
	CC	0.52	ξ_1	0.75	5
			ξ_2	0.75	5
	LD	0.37	ξ_1	0.75	3
			ξ_2	0.75	5

[a] MW, molecular weight; HBA, hydrogen-bond acceptor; HBD, hydrogen-bond donor; RTB, number of rotatable bonds; PSA, polar surface area; CAC, commercially available compounds; CC, clinical candidates; LD, launched drugs.

Estimated Distribution. As mentioned above, both the shape of distributions of physicochemical properties and the correlation between them contain valuable information regarding druglikeness. However, several factors hamper researchers from visualizing the entire picture and effectively utilizing such data. To aid future investigators in capturing the entire picture, we introduced the Nagahara method, originally developed by one of the authors. The estimated distribution of log *P* (*x*-axis) and MW (*y*-axis) for CAC, CC, and LD using the Nagahara method is shown in Figure 3. The probability distribution of CAC is clearly much narrower than the distributions of LD and CC. Such a difference arises from two main factors: the correlation between log *P* and MW, and the fat-tailed distribution. As mentioned in the previous section, a significant correlation can be observed between log *P* and MW in CAC (correlation coefficient 0.39), whereas a weak correlation is observed between these parameters in LD and CC (correlation coefficients 0.17 and 0.28, respectively). Strong correlation is known to cause limited distribution. log *P* and MW distributions are normal in CAC and fat-tailed in CC and LD, and thus the distribution of CAC tends to be limited.

The estimated distributions of HBD (*x*-axis) and PSA (*y*-axis) for CAC, CC, and LD are shown in Figure 4. The probability distribution of CAC clearly differs from those of LD and CC, with the former showing a stubby distribution and the latter two showing elongated distributions. This difference arises from the difference in correlation between HBD and PSA. We also note that the probability distributions of LD and CC is narrower than that of CAC, potentially due to non-normality.

To our knowledge, the Nagahara method is the first and only practical method of generating multivariables based on non-normal distribution free from analytical restrictions while using the parametric distribution system. The method proposed in the present study provides us with deeper insight into the physicochemical properties of a chemical library.

Druglike Score of a Library Based on Correlation Matrix. Table 3 summarizes the druglike score of library (DSL) for several chemical libraries. As described above, the correlation coefficients of LD more closely resemble those of CC than CAC. The DSLs for CC and LD are 0.08 and 0.20, respectively. Surprisingly, the DSL for AnalytiCon is 0.09, which is quite lower than those of other suppliers. This remarkably low value can be explained by the fact that the AnalytiCon library mainly consists of natural derivatives. By 1990, about 80% of all drugs were either natural products

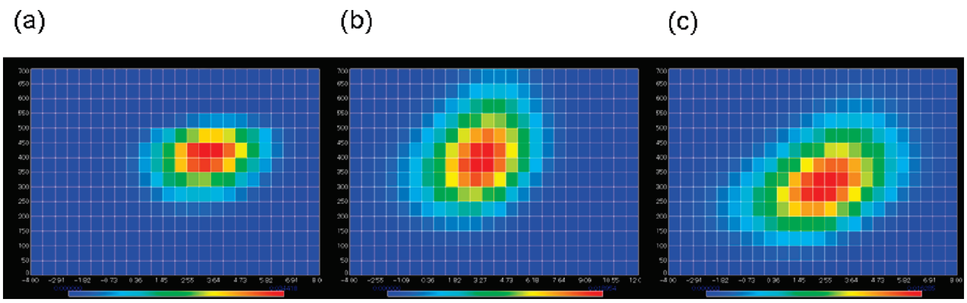


Figure 3. Estimated probability distribution of log *P* (*x*-axis) and MW (*y*-axis): (a) CAC, (b) CC, (c) LD.

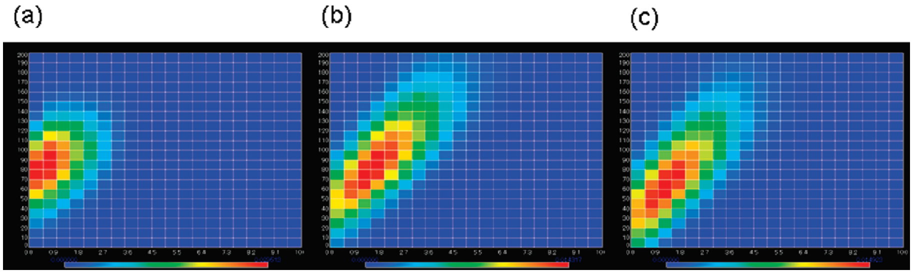


Figure 4. Estimated probability distribution of HBD (*x*-axis) and PSA (*y*-axis): (a) CAC, (b) CC, (c) LD.

Table 3. Druglike Score of Library for Several Chemical Libraries[a]

database	number	DLS
CC	44 843	0.08
CAC	44 140	0.20
AnalytiCon[b]	27 376	0.09
IBS[b]	425 148	0.20
Enamine[b]	1 316 159	0.21
ChemDiv[b]	662 630	0.22
ASINEX[b]	399 583	0.23
Vitas-M[b]	487 261	0.24
Maybridge[b]	56 824	0.25
ChemBridge[b]	422 087	0.25
Bionet[b]	42 33	0.26

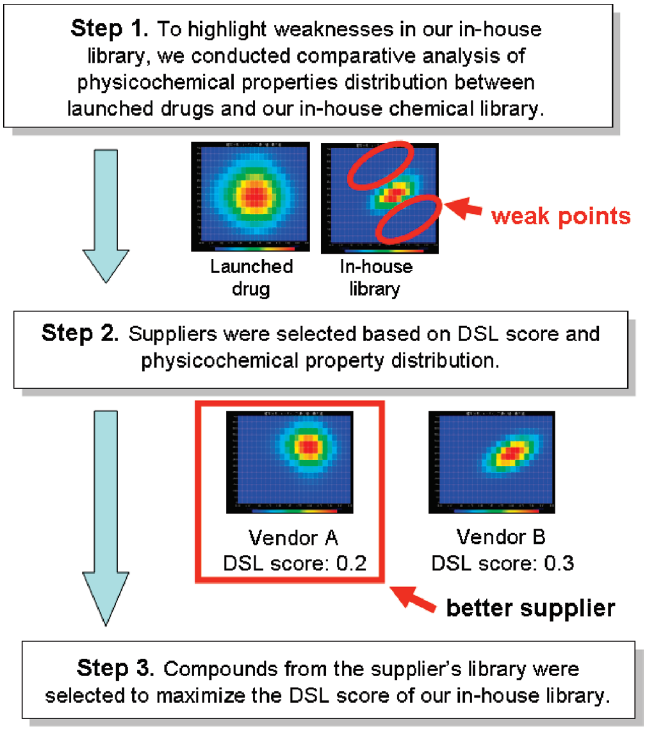
[a] CC, clinical candidates; CAC, commercially available compounds; DLS, druglike score of library. [b] Names of compound suppliers. Chemical structures were obtained from the NAMIKI database (version 200910), which contains 4 458 396 compounds, and Summit Pharmaceuticals database (version 2009 part 3), which contain 5 794 843 compounds. These databases can be obtained from Namiki Shoji Co., Ltd. and Summit Pharmaceuticals International Corp.

or derivatives.[25] While the expansion of synthetic chemistry in the 1990s did cause a notable reduction in the proportion of new natural-product-derived drugs (down to 50%), 13 were still approved in the United States between 2005 and 2007. The similarities between the correlation matrices of LD and AnalytiCon are therefore not so surprising. Similarly, the DSLs for IBS and Enamine are also lower than scores for the other suppliers. A library's DSL may represent a good index for library design, and thus, the simple correlation analysis proposed in the present study may aid computational chemists in evaluating chemical libraries.

Of particular note is the fact that the concept of DSL score differs substantially from the traditional concept as defined by the Lipinski rules. Baurin et al. analyzed the druglike properties for compounds from 23 suppliers,[26] with results showing that 75% of compounds satisfy three Lipinski rules, the second lowest percentage of all suppliers examined. As such, the AnalytiCon library may not be ideal as an HTS library. However, the percentage of LD and CC compounds satisfying these rules was also not high [10 913 of 12 722 compounds (86%)], a value which is lower than that of nearly all suppliers. To account for both launched drugs and clinical candidates, those compounds not satisfying the Lipinski rules should be selected for an HTS library. In this sense, the AnalytiCon library is a good choice for an HTS library, as its physicochemical property distribution is thought to be wide and similar to that of launched drugs (very low DSL score).

While the percentage of compounds satisfying the Lipinski rules is obviously an important and useful indicator in evaluating the quality of a chemical library, it is not the sole, perfect indicator. We must be careful of "Lipinski rule bias". When developing an HTS library, computational chemists usually use Lipinski rule filtering, and as a result, suppliers also tend to use the Lipinski rules when designing new compounds. As shown in Figure 3a, a cliff can be noted at MW = 500 g/mol. For the CAC, quite a greater number of compounds have MWs of 450−500 than 500−550 g/mol. In contrast, for the CC and LD, no clear cliff is noted (Figure 3b,c). Given these observations, compounds with molecular weight slightly exceeding 500 g/mol are thought to be underestimated. Combined use of DSL score with data regarding percentage of compounds satisfying the Lipinski rules may help improve HTS library quality.

Application of the DSL Score and the Nagahara Method to Improving HTS Libraries. Our analysis showed that the probability distribution of CAC is substantially narrower than those of LD or CC (Figure 3). In general, HTS libraries comprised of commercially available compounds are predicted have probability distributions much narrower than those derived from LD or CC databases. In addition, using Lipinski rule filtering further narrows the probability distribution. Compound selection using DSL score and the Nagahara method is one potential solution to this problem of narrow distribution (Figure 5). First, to highlight weak points in the HTS library, differences in the distribution of physicochemical properties between the HTS library and launched drugs are assessed using the Nagahara method,


Step 1. To highlight weaknesses in our in-house library, we conducted comparative analysis of physicochemical properties distribution between launched drugs and our in-house chemical library.

Step 2. Suppliers were selected based on DSL score and physicochemical property distribution.

Step 3. Compounds from the supplier's library were selected to maximize the DSL score of our in-house library.

Figure 5. Schematic representation of the compound selection strategy using DSL score and the Nagahara method.

which can provide a clear, easy-to-understand, visual representation and thereby facilitate communication and good decision-making among researchers in various fields. Next, several suppliers are selected, as bulk buying can reduce compound cost. The most suitable suppliers are then chosen from these initial candidates by assessing DSL score and physicochemical property distributions. Those suppliers with libraries capable of covering the weak points in the existing HTS library are then selected, and finally, compounds are selected to maximize the DSL score for HTS library. As a result, the distribution of physicochemical properties for the in-house library widens, coming to resemble that of launched drugs. Obviously, combined use of this strategy with a conventional approach such as Lipinski rule filtering is also thought to be promising.

ACKNOWLEDGMENT

The authors thank Dr. Mitsuaki Ohta, Dr. Makoto Takeuchi, Dr. Yuzo Matsumoto, Dr. Tatsuya Niimi, Ms. Naoko Katayama, Ms. Ayako Moritomo, Dr. Makoto Oku, Dr. Kenichi Mori, and Dr. Hideyoshi Fuji for their helpful advice.

REFERENCES AND NOTES

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

(2) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615–2623.

(3) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'Rule of Three' for Fragment-Based Lead Discovery. *Drug Discovery Today* **2003**, *8*, 876–877.

(4) Orita, M.; Ohno, K.; Niimi, T. Two 'Golden Ratio' Indices in Fragment-Based Drug Discovery. *Drug Discovery Today* **2009**, *14*, 321–328.

(5) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308–1315.

(6) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The Design of Leadlike Combinatorial Libraries. *Angew. Chem., Int. Ed.* **1999**, *38*, 3743–3748.

(7) Oprea, T. I.; Allu, T. K.; Fara, D. C.; Rad, R. F.; Ostopovici, L.; Bologa, C. G. Lead-like, drug-like or "Pub-like": How Different Are They. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 113–119.

(8) Nagahara, Y. Cross-Sectional-Skew-Dependent Distribution Models for Industry Returns in Japanese Stock Market. *Financ. Eng. Jpn. Mark.* **1995**, *2*, 139–154.

(9) Nagahara, Y. Non-Gaussian Distribution for Stock Returns and Related Stochastic Differential Equation. *Financ. Eng. Jpn. Mark.* **1996**, *3*, 121–149.

(10) Nagahara, Y. The PDF and CF of Pearson Type IV Distributions and the ML Estimation of the Parameters. *Stat. Probab. Lett.* **1999**, *43*, 251–264.

(11) Nagahara, Y. Non-Gaussian Filter and Smoother Based on the Pearson Distribution System. *J. Time Ser. Anal.* **2003**, *24*, 721–738.

(12) Nagahara, Y. A Method of Calculating the Downside Risk by Multivariate Non-Normal Distributions. *Asia−Pac. Financ. Mark.* **2008**, *15*, 175–184.

(13) Nagahara, Y.; Kitagawa, G. Non-Gaussian Stochastic Volatility Model. *J. Comput. Finance* **1999**, *2*, 33–47.

(14) Kato, H.; Nagahara, Y.; Arai, S.; Sawada, H.; Makino, S. Frequency Domain Pearson Distribution Approach for Independent Component Analysis (FD-Pearson-ICA) Blind Source Separation. *IEEE Trans. Speech Aud. Proc.* **2009**, *17*, 639–649.

(15) Tantar, A.; Melab, N.; Talbi, E. An Analysis of Dynamic Mutation Operators for Conformational Sampling, *Biologically-Inspired Optimisation Methods*; Lewis, A., Mostaghim, S., Randall, M., Eds.; Springer: Berlin, 2009; Vol. 210, pp 291−323,

(16) Pearson, K. Memoir on Skew Variation in Homogeneous Material. *Philos. Trans. R. Soc. A* **1895**, *186*, 343–414.

(17) Elderton, W. P.; Johnson, N. L. *Systems of Frequency Curves*; Cambridge University Press: London, 1969.

(18) Johnson, N. L., Kotz, S. Balakrishnan, N. *Continuous Univariate Distributions—1*, 2nd ed.; John Wiley: Chichester, UK, 1994.

(19) Stuart, A., Ord, J. K. *Kendall's Advanced Theory of Statistics, Vol. 1, Distribution Theory*, 6th ed.; John Wiley: Chichester, UK, 1994.

(20) Nagahara, Y. A Method of Simulating Multivariate Non-Normal Distributions by the Pearson Distribution System. *Comput. Stat. Data Anal.* **2004**, *47*, 1–29.

(21) Matsumoto, M.; Nishimura, T. Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator. *ACM Trans. Model. Comput. Simul.* **1998**, *8*, 3–30.

(22) Irwin, J. J.; Shoichet, B. K. ZINC—A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model* **2005**, *45*, 177–182.

(23) Contract Research Organisation India, Indian CROs GVK BIO. http://www.gvkbio.com/ (accessed Nov 4, 2009).

(24) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragment Methods: An Analysis of ALogP and CLogP Methods. *J. Phys. Chem. A* **1998**, *102*, 3762–3772.

(25) Li, J. W.-H.; Vederas, J. C. Drug Discovery and Natural Products: End of an Era or an Endless Frontier. *Science* **2009**, *325*, 161–165.

(26) Greaney, P.; Morley, D.; Hubbard, R. E. Drug-like Annotation and Duplicate Analysis of a 23-Supplier Chemical Database Totalling 2.7 Million Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 643–651.

CI100023S