Do Biologically Relevant Compounds Have More Chance To Be Drugs?

De-Xin Kong,*[,†,‡] Wei Ren,[‡] Wei Lü,[‡] and Hong-Yu Zhang[†]

College of Life Science and Technology, Huazhong Agricultural University, Wuhan 430070, China, and Shandong Provincial Research Center for Bioinformatics Engineering and Technique, Center for Advanced Study, Shandong University of Technology, Zibo 255049, China

Received July 1, 2009

To prove the innate advantages of endogenous compounds/fragments for drug discovery and development, a novel index termed biological relevance (*BR*) is proposed. The results clearly indicate its ability to distinguish between synthetic chemicals, bioactive compounds, drug candidates, and launched drugs. Primarily, the average *BR* of the databases investigated decreases in the order DNP > CMC > ACD-3D > MDDR. Second, for compounds with the same bioactivity, drugs (CMC) possess higher average *BR* than their candidates (MDDR). These results suggest that compounds with higher *BR* have more chance to survive the drug development pipeline. Third, the above conclusion is supported by the fact that compounds in the later development phases possess higher *BR* than those in the earlier phases. Comparisons were made between *BR* and other indices, including toxicity, druglikeness, and natural productlikeness.

INTRODUCTION

Despite its history of success, modern pharmaceutical industry is facing the challenge of "more investment - less outcome". Global expenditure on research has doubled since 1991, but the number of new entity drugs approved annually has fallen by 50% or even more.[1,2] To change this situation, pharmaceutical companies are attempting to alter their paradigms from random screening to rational design.

Experimentally, high-content screening,[3,4] systems biology,[5] fragment-based design,[6] and multitarget (or synergistic, multicomponent) drugs[7,8] have been proposed to avoid the shortcomings of high-throughput screening (HTS) and to change the current status of one drug - one target. Also, pharmaceutical companies shifted their interest back to natural products (NPs),[9−12] which proved to be an effective approach.[13,14]

Theoretically, a series of indices, including molecular diversity,[15] druglikeness,[16,17] druggability,[18−21] leadlikeness,[22,23] predicted ADMET properties,[24] etc. were proposed to remove the redundancy and improve the quality of compound libraries. Among these, druglikeness was the most widely accepted and used. However, druglikeness was proposed on the basis of known oral drugs, which limited its application in searching for new drug entities.[16,25]

The term "biological relevance" or "biologically relevant" has been used in several reports.[26−32] However, the definition, calculation algorithm of this term, and whether biologically relevant compounds have more chance to survive the drug development pipeline remain to be investigated.

To address the above issues, we here provided our algorithm of a novel index, named "biological relevance (*BR*)". First, a standard biologically relevant database, named Biorelevant Representative Compounds Database (BRCD), was constructed, which includes 2000 compounds diversely selected from KEGG ligands.[33,34] Then, the biological relevance (*BR*) of a compound or a database was defined on the basis of Tanimoto[35] similarity calculation (see Materials and Methods for further details).

THE ALGORITHM

Biological relevance is a measurement of the possibility of biological origination for a compound or a database. It can be evaluated as the similarity (*S*) between the objective compound (Q) and standard biologically relevant compounds. Since there are 2000 compounds in BRCD, the similarity is characterized as a 2000-dimensional array. The *BR* score can be assigned simply to the largest similarity value, as an example, with compound A

$$BR_A = S_{QA} \quad (1)$$

Resembling the set-theoretic notation of tanimoto similarity, the biologically relevant chemical space (fragments or fingerprint bits) occupied by Q can be denoted as $Q \cap A$.

However, BRCD were diversely selected, considering only the largest similarity is biased against compounds that are similar to more than one BRCD compound (e.g., the second and the third compound in the Supporting Information (SI), Table S1). Thus, the secondary largest similarity (as an example, with compound B) should be considered. In this situation, the biologically relevant chemical space occupied by Q can be denoted as

$$(Q \cap A) \cup (Q \cap B) = (Q \cap A) + (Q \cap B) - (Q \cap A \cap B) = (Q \cap A) + (Q \cap B) - (Q \cap A) \cap (Q \cap B) \quad (2)$$

Therefore, the *BR* score can be calculated as

$$BR_{AB} = S_{QA} + S_{QB} - S_{QA}S_{QB} = S_{QA} + (1 - S_{QA})S_{QB} = BR_A + (1 - BR_A)S_{QB} \quad (3)$$

where A and B are BRCD compounds. The purpose to minus $S_{QA}S_{QB}$ is to remove the probability of chemical space that

* Corresponding author phone and fax: +86-27-8728 0877; e-mail: dxkong@mail.hzau.edu.cn.
† Huazhong Agricultural University.
‡ Shandong University of Technology.

10.1021/ci900229c CCC: $40.75
Published on Web 09/24/2009

Table 1. Average *BR* Scores of the 4 Databases (7939 Random Compounds) with Depths 1−5

database	BR_1	BR_2	BR_3	BR_4	BR_5
DNP	0.408	0.579	0.682	0.750	0.799
CMC	0.310	0.457	0.555	0.626	0.680
ACD-3D	0.252	0.396	0.497	0.572	0.631
MDDR	0.223	0.362	0.463	0.543	0.606

overlapped by all three compounds Q, A, and B. This part is counted repeatedly and should be subtracted from $S_{QA}+S_{QB}$. Since compounds A and B are independent (diversely selected), the probability should be equal to $S_{QA}S_{QB}$.

Similarly, we can include the third largest similarity value. Because

$$(Q \cap A) \cup (Q \cap B) \cup (Q \cap C) = (Q \cap A) + (Q \cap B) + (Q \cap C) - (Q \cap A \cap B) - (Q \cap A \cap C) - (Q \cap B \cap C) + (Q \cap A \cap B \cap C) \quad (4)$$

BR can be calculated as

$$BR_{ABC} = S_{QA} + S_{QB} + S_{Qc} - S_{QA}S_{QB} - S_{QA}S_{QC} - S_{QB}S_{QC} + S_{QA}S_{QB}S_{QC} = BR_{AB} + (1 - BR_{AB})S_{QC} \quad (5)$$

Thus, the universal formula for *BR* will be

$$\begin{gathered} BR_1 = S_1 \\ BR_2 = S_1 + S_2 - S_1 \times S_2 = BR_1 + (1 - BR_1) \times S_1 \\ BR_3 = S_1 + S_2 + S_3 - S_1 \times S_2 - S_2 \times S_3 - S_1 \times S_3 + S_1 \times S_2 \times S_3 = BR_2 + [1 - BR_2] \times S_3 \\ \cdots \\ BR_d = BR_{d-1} + (1 - BR_{d-1}) \times S_d \end{gathered} \quad (6)$$

where S_d represents the dth largest similarity between the objective molecule and BRCD compounds; therefore, d represents the calculation depth. The calculated *BR* score will increase with increased depth (d); therefore, comparison should be performed at the same depth.

To determine an appropriate depth, 7939 compounds selected at random from the four databases investigated (DNP,[36] CMC, ACD-3D, and MDDR,[37] see Materials and Methods for detailed introductions) were analyzed. As the results show (Table 1), whatever depth was adopted, the average *BR* of the databases always decreased in the order DNP > CMC > ACD-3D > MDDR. The correlation coefficients between BR_1-BR_4 and BR_5 were 0.844, 0.940, 0.979, and 0.996, respectively (calculated with 8000 random compounds). The default depth was set to 3 for discriminant efficiency and better distribution (SI, Figure S1).

The *BR* score is restricted between 0 and 1. A score of 0 means that this compound is not biologically relevant in any sense; i.e., the query compound is quite different from all BRCD compounds. A score of 1 denotes the other extreme situation, in which there is a compound in BRCD with the same structure as the query compound. The *BR* score of a database can be calculated as the arithmetic mean of the *BR* scores of its member compounds, which will be also restricted between 0 and 1.

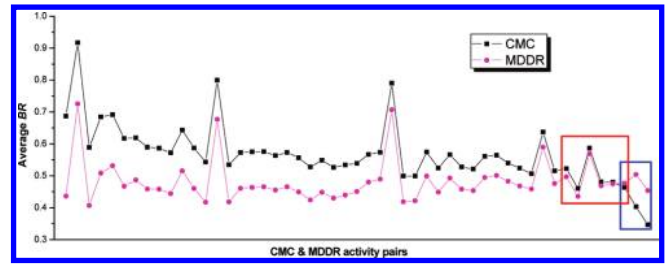

Figure 1. Comparison of the activity pairs in MDDR and CMC. For compounds with the same bioactivity, the average *BR* of CMC compounds is higher than that of MDDR compounds (48/51 activity pairs, see SI, Table S2 for detailed data). The exceptions are marked in the blue rectangle, with activities of antiviral, antiprotozoal, and antiAIDS. A total of 43 pairs were significantly different (on the left of the red rectangle). Therefore, a conclusion can be drawn that drugs generally possess higher *BR* scores than their candidates, which indicates the importance of the *BR* score in drug discovery. From left to right, the activities are cardiotonic, androgenic, anticoagulant, antipsoriatic, dermatologic, hypolipidemic, antiinflammatory, antiseptic, antiarthritic, anticonvulsant, antibacterial, antiallergic, vasodilator, estrogenic, antiemetic, antiulcerative, mucolytic, anticholinergic, antiparkinsonian, analgesic, bronchodilator, antimigraine, antiarrhythmic, antidepressant, antihistamine, antianginal, stimulant, antineoplastic, progestin, anxiolytic, antisecretory, antitussive, antihypertensive, antifungal, antiasthmatic, 5-HT receptor antagonist, muscle relaxant, Ca channel antagonist, antimalarial, platelet aggregation inhibitor, antipsychotic, antibiotic, immunomodulator, anthelmintic, hypnotic, anesthetic, antidiabetic, diuretic, antiviral, antiprotozoal, and antiAIDS.

RESULTS

The Rationality of BR. As mentioned above, the *BR* scores of the investigated databases decrease significantly in the order DNP > CMC > ACD-3D > MDDR (Table 1). Set 0.5 as a critical value, 81.33% of DNP compounds, 58.45% of CMC compounds, 46.68% of ACD-3D compounds, and 31.68% of MDDR compounds are biologically relevant.

The DNP compounds were isolated from natural organisms; therefore, they should have very high *BR* scores. Compounds in CMC have been used as medicinal agents, and some of them are NPs or NP derivatives. Therefore, the *BR* score of CMC compounds should be rather high but lower than those of DNP compounds. ACD-3D compounds have lower *BR* scores because most of them are synthetic. Therefore, the origination of these databases and their *BR* scores are consistent.

Both CMC and MDDR compounds possess biological activity. Usually both are treated as druglike compounds,[38,39] and their *BR* scores should be comparable. Surprisingly, the *BR* of MDDR compounds is the lowest in all the databases. It would be very interesting to affirm this unexpected result. Therefore, 51 pairs of compound sets were extracted from MDDR and CMC according to their annotated activity.

For compounds with the same activity, the average *BR* of the CMC subset is always higher than that of MDDR (Figure 1; SI, Table S2), except for the activities of antiviral, antiprotozoal, and antiAIDS. Comparison of means using Student's *t*-test showed that most pairs (43/48) were significantly different. Since most MDDR compounds are bioactive but failed to be drugs, it can be considered that drugs generally possess higher *BR* scores than their candidates.

If the conclusion mentioned above is right, compounds at later development phases should possess higher *BR* scores than those at earlier phases. Thus, compounds in MDDR

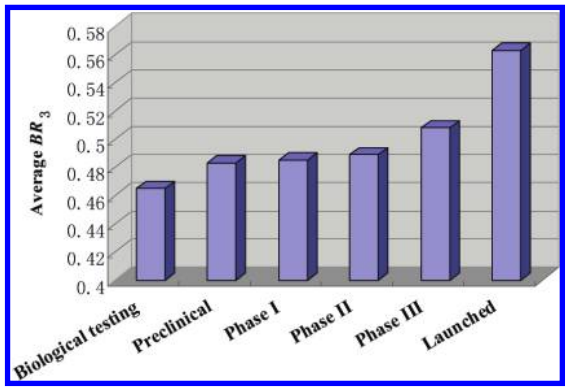

Figure 2. Compounds at later development phases possess higher biological relevance (*BR)* than former phases. The result suggests that compounds with higher *BR* have more chance to be developed as drugs.

were extracted, grouped, and analyzed according to their development phase. Interestingly, the average *BR* scores for compounds of ordinal phases do increase (Figure 2 and SI, Table S3). This result confirmed the measurability of *BR* and indicated the significance of *BR* in drug development; i.e., compounds with higher *BR* scores have more chance of surviving the drug discovery and development pipeline.

BR Based on Primary Metabolites. Because the KEGG compounds are a mixture of various sources, it may be argued that the results were influenced by the drug metabolites in KEGG.

To clarify this suspicion, similar calculations were performed with Pri-BRCD, a database similar to BRCD but constructed with primary metabolites (1000 diverse compounds extracted from a total of 2831 KEGG primary metabolites). The results based on primary metabolites (left in Figure 3a,c and top in Figure 3b) were quite similar to those based on BRCD. Because the number of primary metabolites was limited, the discriminant power based on Pri-BRCD is a little weaker than that based on BRCD. Definitely, there are no (or very limited) drug metabolites in primary metabolites.

On the other hand, there are a lot of drug candidates in MDDR. If drug metabolites in KEGG can increase the *BR* score of drug compounds (CMC), then they should increase the *BR* score of MDDR at the same time, whereas the *BR* score of MDDR was the lowest in all the databases. Therefore, the calculation of *BR* was not influenced by possible drug metabolites in KEGG.

BR Based on Natural Products. Natural products have been optimized in a long-term natural selection process for optimal interaction with biological macromolecules. Therefore, NPs were considered as a valuable source for drug leads. Recently, based on Bayesian discriminant analysis of DNP and Novartis' in-house collection of synthetic compounds, Ertl et al. presented an algorithm of natural productlikeness (NP-likeness).[40] Property distribution and common fragments of NPs were also identified through chemoinformatics analysis.[41−43]

To compare the significance of NPs in terms of deducing an index for library design, a similar study was also performed with a database of NPs (2000 diverse compounds selected from >160,000 DNP compounds, named DNP-BRCD). Although similar results were obtained (right in Figure 3a,c and bottom in Figure 3b), the discriminant power

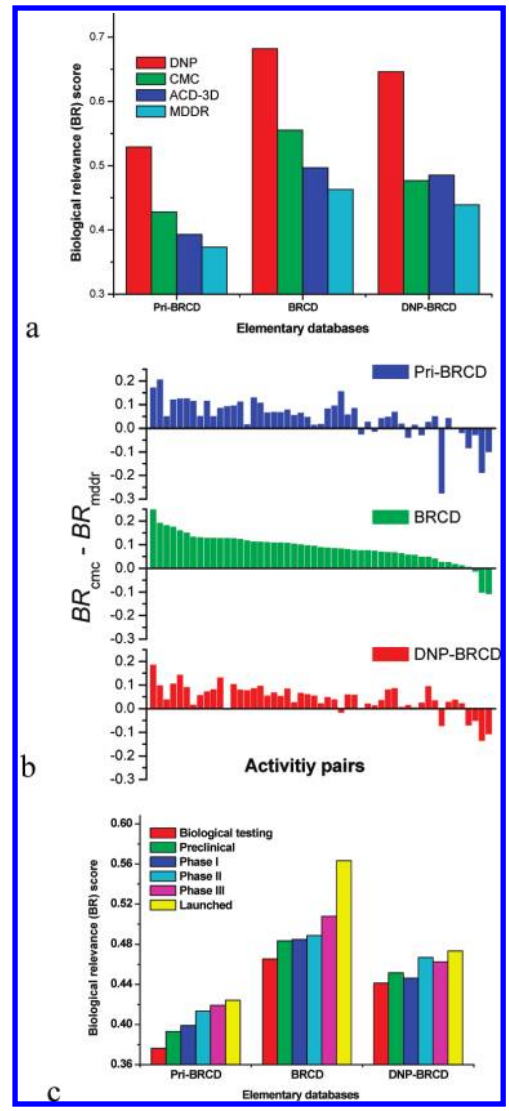

Figure 3. Comparison of the discriminant efficiency based on different elementary databases (Pri-BRCD, BRCD, and DNP-BRCD). The results were obtained with the same protocol proposed in this paper: a) to distinguish different databases, b) to compare MDDR and CMC compounds with the same bioactivity. The order of activities is the same as that in Figure 1, and c) comparison of compounds at different development phases. As the results show, the score based on DNP-BRCD has the weakest discriminant power.

of DNP-BRCD is the weakest among the three elementary databases (Pri-BRCD, BRCD, DNP-BRCD).

DNP includes all kinds of known natural compounds with various kinds of functions and activities. Why the discriminant power of the index deduced from NPs is weaker than that deduced from KEGG compounds or primary metabolites? In our opinion, the NPs' diversity (both the structure and the bioactivity) is too expansive to be used for general purpose in drug discovery. For example toxic or noxious NPs should be excluded for library design or the deduction of an index.

Index based on NPs has weaker discriminate power than based on KEGG compounds (BRCD). Nevertheless, it should be noticed that NPs are very important and instructive. Especially, in Figure 3b, the difference of the three elementary databases can give valuable clues for related drug design and need further analysis.

Table 2. Comparison of the *BR* Score and Lipinski Violations with 8000 Random Compounds

Lipinski violation	average *BR*	standard deviation	number of compounds
0	0.515	0.205	4692
1	0.587	0.222	1475
2	0.601	0.217	969
3	0.729	0.219	805
4	0.709	0.190	59

Comparison of BR and Toxicity. Why do compounds with low *BR* scores hardly become drugs? The first interpretation that sprang to our mind was their toxicity. To test this assumption, 59 groups of acute toxicity data sets with different testing species, end points and routines of administration were extracted from the MDL Toxicity-Finder database. However, no direct correlation was observed between *BR* and toxicity (SI, Table S4). This result is consistent with early reports that "natural does not mean harmless".[44,45] Thus, biologically relevant compounds must possess other favorable advantages, perhaps better ADME properties; however, these are difficult to characterize using simple descriptors. Therefore, comparison between *BR* and druglikeness was made.

Comparison of BR and Druglikeness. The average Lipinski's rules-of-five (*Ro5*)[17] violations of DNP, CMC, ACD-3D, and MDDR were 0.803, 0.313, 0.280, and 0.693, respectively. Combining the results listed in Table 1, DNP has the lowest level of druglikeness but the highest *BR* scores. ACD-3D has the highest level of druglikeness, which reflects the influence of the concept druglikeness in current drug discovery, but rather low *BR* scores.

Calculations of 8000 random compounds extracted from the CMC, ACD-3D, MDDR, and DNP databases showed that the *BR* score increases with the increase of Lipinski violations (druglikeness decreased; Table 2). Therefore, to some extent, *BR* is different from druglikeness. Since Lipinski's *Ro5* does not include information for natural products,[32,46] this result is not unexpected.

Comparing with druglikeness, *BR* was derived totally from KEGG metabolites (no information for current drugs was considered), which makes this index escape the restriction of current approved drugs and can be used for future drug discovery.

Properties That Are Correlated with BR. To characterize the chemical space of *BR* compounds, >80 molecular descriptors were calculated for the 8000 random compounds mentioned above. Analysis showed weak but statistically significant correlations between *BR* and the number of chiral centers, rotatable bonds, NSOP atoms, secondary alcohols and hydroxyl groups, etc. (SI, Table S5). Therefore, the factors that influence *BR* are multiple and need further study.

DISCUSSION

Current drug discovery depends largely on random screening, either high-throughout screening (HTS) in vitro or virtual screening (VS) in silicon. Because the number of compounds that can be available or can be formed theoretically is very huge,[47,48] several indexes (i.e., druglikeness, leadlikeness, NP-likeness) were proposed to reduce the number of compounds that need to be synthesized or to be screened. Most of these indexes are restricted by current knowledge/drugs, and no information about protein structure is included. Directly using of protein structure information is restricted by protein structure determination and protein−ligand docking technologies. In fact, the information about protein pockets is encoded in the structures of its substrates,[49] i.e., the endogeneous compounds/fragments. Considering that the amount of protein folds is finite,[50−53] the structural categories of the ligands that can be bound to proteins should be limited too. That is to say, for the molecules that can be formed theoretically, only some of them can bind to proteins and displays biological functions. Therefore, we speculate that it may be possible to use the chemical space of the substrates as guidance for library design. Only in this way, could we avoid the limitation of our current knowledge and use the information about protein pockets indirectly.

As the results of this study shown, compounds with higher *BR* scores have more chance of surviving the development pipeline. Therefore, *BR* can be used as an index of druggability, which will be useful in library optimization, initial filtering for virtual screening, or the construction of virtual combinatorial libraries. A high-quality compounds library for HTS/VS should possess maximum diversity, druglikeness, etc., and possess a high degree of biological relevance.

Due to the superiority to bind with transport carriers, naturally occurring metabolites offer greater bioavailability, better distribution, and a greater chance of bioactivity discovery.[54,55] Therefore, "metabolitelikeness" can be used as a criterion for library design.[56,57] The metabolites should possess high biological relevance, which gives another explanation for the advantage of biological relevance.

The algorithm described in this manuscript provides an effective approach for library comparison and index derivation. Compared with other methods, no information of negative library (e.g., synthetic compounds database) is needed. The value of the deduced index can be restricted to between 0 and 1, which makes them more comparable than nonscaled indices (e.g., Bayesian discriminant score). The similarities between a candidate and 2000 BRCD compounds constitute a 2000-dimension descriptor, which can be used for various purposes, such as predicting the bioactivity class of a compound or some ADME properties.

Although structurally similar compounds do not always show similar biological activity,[58,59] the present study displayed distinct application potential of *BR*. It should be noted that MolPrint 2D[60,61] is a rather simple method for similarity calculation. If a more sophisticated similarity protocol can be employed during the BRCD compounds sampling and *BR* calculation, the results will be more significant.

MATERIALS AND METHODS

Construction of the Biorelevant Representative Compound Database. KEGG ligands collected most known compounds that existed in biological metabolism pathways.[33,34] These ligands can be treated as biologically relevant compounds and should possess the potential to bind with proteins. Our data (KEGG compounds subgroup, 12127 in total) were downloaded from the KEGG ftp server on Sep. 6, 2006. Additional refinement was processed as follows.

1. For multicomponent records, small fragments (counterions in salts, solvent molecules) were removed, and only the largest fragment was reserved.

2. Hydrogen was added to fulfill the valences of heavy atoms and to neutralize the molecular charge.

3. Inorganic compounds, small molecules (molecular mass <100 Da), and very large molecules with more than 500 atoms were removed.

4. Compounds with the same structure as known drugs (CMC, KEGG drugs) were discarded.

5. The database was reduced to 2000 molecules based on diversity for the reasons that a) minimizing the number of compounds can make computation more convenient and b) there were many similar or clustered compounds in the original database. Only one compound of each group was reserved as a representative.

These 2000 compounds formed the Biorelevant Representative Compound Database (BRCD), which can be thought of as a diverse representative of the endogenous compounds. Since the selection was diversity-based, BRCD is a space uniform-distribution sampling instead of a probability representation.

The primary metabolites were extracted from the KEGG primary metabolism pathways. Before the construction of Pri-BRCD, compounds that duplicated with KEGG phytochemical compounds were removed.

The Test Sets. Random sample of four databases, the Comprehensive Medicinal Chemistry (CMC), the MDL Drug Data Report (MDDR), the Available Chemicals Directory with 3D structure (ACD-3D), and the Dictionary of Natural Products (DNP) were used to testify the efficiency of our algorithm. The first three databases (CMC, MDDR, ACD-3D) and toxic data (ToxicFinder database) were provided by MDL Information Systems Inc.[37] DNP was provided by CRC Press.[36] In this paper, CMC was selected to represent approved drugs, MDDR represents bioactive compounds that failed launching as drugs, ACD-3D represents synthetic compounds, and DNP represents natural products.

Calculations. KEGG reduction for BRCD construction was performed with the library analysis module in Cerius 2 software.[62] Monte Carlo diverse selection method was used, since the analysis of self-similarity matrixes of reduced database showed that this method could give a more diverse result than cluster-based method.

Tanimoto[35] similarity calculation was performed with a Perl program rewritten on the basis of MolPrint 2D,[60,61] and all other noncommercial data are available from the authors upon request. A Tanimoto similarity is defined as follows

$$S_T = \frac{N_c}{N_a + N_b - N_c}$$

where N_a and N_b are the number of bits set for fingerprints of molecules A and B, respectively, and N_c is the bits set present in molecules A and B.

Batch processing was done with Pipeline Pilot[63] or UNIX shell scripts. Statistical analysis was done with R statistics packages.[64] Molecular descriptors were calculated with Sybyl,[65] Pipeline Pilot, Cerius 2, and Dragon.[66]

ACKNOWLEDGMENT

This study was supported by the National Basic Research Program of China (grant 2010CB126100) and the National Natural Science Foundation of China (grant 30870520). We are grateful to Dr. Mingyue Zheng, Ms. Xue-Juan Li, and Dr. Ping Wu for critically reading the manuscript.

Supporting Information Available: Distribution curve of *BR* scores with calculation depth form 1 to 5 (Figure S1), calculation examples of *BR* score (Table S1), average BR_3 of CMC and MDDR compounds with the same bioactivity (Table S2), average *BR* scores of compounds at development phases (Table S3), the correlation analysis between BR_3 and toxicity (Table S4), the correlation coefficient between BR_3 and calculated descriptors (Table S5), and a zipped sd file for BRCD database. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Buehler, L. K. Advancing Drug Discovery - Beyond Design. *Pharmagenomics* **2004**, *4*, 24–26.
(2) Hughes, B. 2007 FDA drug approvals: a year of flux. *Nat. Rev. Drug Discovery* **2008**, *7*, 107–109.
(3) Haney, S. A.; LaPan, P.; Pan, J.; Zhang, J. High-content screening moves to the front of the line. *Drug Discovery Today* **2006**, *11*, 889–894.
(4) Liptrot, C. High content screening - from cells to data to knowledge. *Drug Discovery Today* **2001**, *6*, 832–834.
(5) Materi, W.; Wishart, D. S. Computational systems biology in drug discovery and development: methods and applications. *Drug Discovery Today* **2007**, *12*, 295–303.
(6) Bembenek, S. D.; Tounge, B. A.; Reynolds, C. H. Ligand efficiency and fragment-based drug discovery. *Drug Discovery Today* **2009**, *14*, 278–283.
(7) Keith, C. T.; Borisy, A. A.; Stockwell, B. R. Multicomponent therapeutics for networked systems. *Nat. Rev. Drug Discovery* **2005**, *4*, 71–78.
(8) Zimmermann, G. R.; Lehár, J.; Keith, C. T. Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug Discovery Today* **2007**, *12*, 34–42.
(9) Desai, M. C.; Chackalamannil, S. Rediscovering the role of natural products in drug discovery. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 436–437.
(10) Paterson, I.; Anderson, E. A. The renaissance of natural products as drug candidates. *Science* **2005**, *310*, 451–453.
(11) Newman, D. J.; Cragg, G. M. Natural products as sources of new drugs over the last 25 years. *J. Nat. Prod.* **2007**, *70*, 461–477.
(12) Ojima, I. Modern natural products chemistry and drug discovery. *J. Med. Chem.* **2008**, *51*, 2587–2588.
(13) Balamurugan, R.; Dekker, F. J.; Waldmann, H. Design of compound libraries based on natural product scaffolds and protein structure similarity clustering (PSSC). *Mol. Biosyst.* **2005**, *1*, 36–45.
(14) Breinbauer, R.; Vetter, I. R.; Waldmann, H. From protein domains to drug candidates-natural products as guiding principles in the design and synthesis of compound libraries. *Angew. Chem., Int. Ed.* **2002**, *41*, 2879–2890.
(15) Gorse, A. D. Diversity in medicinal chemistry space. *Curr. Top. Med. Chem.* **2006**, *6*, 3–18.
(16) Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discovery* **2007**, *6*, 881–890.
(17) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
(18) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007**, *25*, 71–75.
(19) Keller, T. H.; Pichota, A.; Yin, Z. A practical view of 'druggability'. *Curr. Opin. Chem. Biol.* **2006**, *10*, 357–361.
(20) Sirois, S.; Hatzakis, G.; Wei, D.; Du, Q.; Chou, K. C. Assessment of chemical libraries for their druggability. *Comput. Biol. Chem.* **2005**, *29*, 55–67.
(21) Sugiyama, Y. Druggability: selecting optimized drug candidates. *Drug Discovery Today* **2005**, *10*, 1577–1579.
(22) Hann, M. M.; Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255–263.
(23) Rishton, G. M. Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discovery Today* **2003**, *8*, 86–96.

(24) Modi, S. Positioning ADMET in silico tools in drug discovery. *Drug Discovery Today* **2004**, *9*, 14–15.
(25) A decade of drug-likeness. *Nat. Rev. Drug Discovery* **2007**, *6*, 853–853 (editorial).
(26) Putta, S.; Landrum, G. A.; Penzotti, J. E. Conformation mining: an algorithm for finding biologically relevant conformations. *J. Med. Chem.* **2005**, *48*, 3313–3318.
(27) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17272–17277.
(28) Koch, M. A.; Waldmann, H. Protein structure similarity clustering and natural product structure as guiding principles in drug discovery. *Drug Discovery Today* **2005**, *10*, 471–483.
(29) Larsson, J.; Gottfries, J.; Muresan, S.; Backlund, A. ChemGPS-NP: tuned for navigation in biologically relevant chemical space. *J. Nat. Prod.* **2007**, *70*, 789–794.
(30) Meggers, E. Exploring biologically relevant chemical space with metal complexes. *Curr. Opin. Chem. Biol.* **2007**, *11*, 287–292.
(31) Shelat, A. A.; Guy, R. K. Scaffold composition and biological relevance of screening libraries. *Nat. Chem. Biol.* **2007**, *3*, 442–446.
(32) Feher, M.; Schmidt, J. M. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218–227.
(33) Kanehisa, M.; Goto, S.; Kawashima, S.; Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **2002**, *30*, 42–46.
(34) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **2006**, *34*, D354–357.
(35) Holliday, J. D.; Ranade, S. S.; Willett, P. A fast algorithm for selecting sets of dissimilar molecules from large chemical databases. *Quant. Struct.-Act. Rel.* **1995**, *14*, 501–506.
(36) *Dictionary of natural products, version 17.1*; Chapman & Hall/CRC: London, England, 2008.
(37) *MDL databases (CMC, ACD-3D, MDDR, ToxFinder), version 2004.1*; Elsevier MDL: San Leandro, CA, 2004.
(38) Ajay, A.; Walters, W. P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules. *J. Med. Chem.* **1998**, *41*, 3314–3324.
(39) Frimurer, T. M.; Bywater, R.; Naerum, L.; Lauritsen, L. N.; Brunak, S. Improving the odds in discriminating "drug-like" from "non drug-like" compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1315–1324.
(40) Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural product-likeness score and its application for prioritization of compound libraries. *J. Chem. Inf. Model.* **2008**, *48*, 68–74.
(41) Ertl, P.; Schuffenhauer, A. Cheminformatics analysis of natural products: lessons from nature inspiring the design of new drugs. *Prog. Drug Res.* **2008**, *66* (217), 219–235.
(42) Grabowski, K.; Schneider, G. Properties and Architecture of drugs and natural products revisited. *Curr. Chem. Biol.* **2007**, *1*, 115–127.
(43) Wetzel, S.; Schuffenhauer, A.; Roggo, S.; Ertl, P.; Waldmann, H. Cheminformatics analysis of natural products and their chemical space. *Chimia* **2007**, *61*, 355–360.
(44) Ernst, E. When natural is not harmless. *Int. J. Clin. Pract.* **2006**, *60*, 380–380.
(45) Natural doesn't mean harmless. *New Sci.* **2008**, *197*, 5–5 (editorial).
(46) Ganesan, A. The impact of natural products upon modern drug discovery. *Curr. Opin. Chem. Biol.* **2008**, *12*, 306–317.
(47) Fink, T.; Reymond, J. L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342–353.
(48) Fink, T.; Bruggesser, H.; Reymond, J. L. Virtual exploration of the small-molecule chemical universe below 160 Daltons. *Angew. Chem., Int. Ed.* **2005**, *44*, 1504–1508.
(49) Ji, H. F.; Kong, D. X.; Shen, L.; Chen, L. L.; Ma, B. G.; Zhang, H. Y. Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery. *Genome Biol.* **2007**, *8*, R176.
(50) Sadreyev, R. I.; Grishin, N. V. Exploring dynamics of protein structure determination and homology-based prediction to estimate the number of superfamilies and folds. *BMC Struct. Biol.* **2006**, *6*, 6.
(51) Leonov, H.; Mitchell, J. S.; Arkin, I. T. Monte Carlo estimation of the number of possible protein folds: effects of sampling bias and folds distributions. *Proteins* **2003**, *51*, 352–359.
(52) Govindarajan, S.; Recabarren, R.; Goldstein, R. A. Estimating the total number of protein folds. *Proteins* **1999**, *35*, 408–414.
(53) Liu, X.; Fan, K.; Wang, W. The number of protein folds and their distribution over families in nature. *Proteins* **2004**, *54*, 491–499.
(54) Dobson, P. D.; Kell, D. B. Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule. *Nat. Rev. Drug Discovery* **2008**, *7*, 205–220.
(55) Kell, D. B. Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug Discovery Today* **2006**, *11*, 1085–1092.
(56) Dobson, P. D.; Patel, Y.; Kell, D. B. 'Metabolite-likeness' as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Discovery Today* **2009**, *14*, 31–40.
(57) Gupta, S.; Aires-de-Sousa, J. Comparing the chemical spaces of metabolites and available chemicals: models of metabolite-likeness. *Mol. Diversity* **2007**, *11*, 23–36.
(58) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods. *Drug Discovery Today* **2002**, *7*, 903–911.
(59) Gozalbes, R. Does structural similarity reflect biological activity. *Drug Discovery Today* **2006**, *11*, 957–957.
(60) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naive Bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.
(61) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.
(62) *Cerius 2, version 4.11L*; Accelrys: San Diego, CA, 2006.
(63) *Pipeline Pilot Student Edition, version 6.1.5*; SciTegic Accelrys: San Diego, CA, 2007.
(64) *R statistics packages, version 2.6.2*; The R Foundation for Statistical Computing: 2008;ISBN 3-900051-07-0.
(65) *Sybyl, version 7.0*; Tripos: Louis, MO, 2004.
(66) *Dragon, version 5.0*; Talete SRL: Milan, Italy, 2005.

CI900229C