JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Definition of Drug-Likeness for Compound Affinity

Yoshifumi Fukunishi[†,‡,*] and Haruki Nakamura[§]

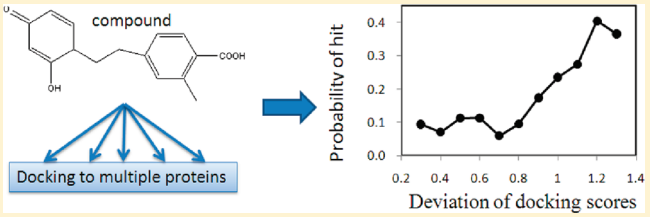
[†]Biomedicinal Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-3-26, Aomi, Koto-ku, Tokyo 135-0064, Japan

[‡]Pharmaceutical Innovation Value Chain, BioGrid Center Kansai, 1-4-2 Shinsenri-Higashimachi, Toyonaka, Osaka 560-0082, Japan

[§]Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

Supporting Information

ABSTRACT: We proposed a new definition of drug-likeness based on protein−compound docking simulation. Active and decoy compounds of 40 target proteins were investigated. These compounds were docked to protein sets consisting of 53−160 proteins. The protein sets did not include the target proteins. The average value and deviation of docking scores against the multiple proteins were calculated for each compound. Our study revealed that the docking scores of active compounds are more widely distributed than those of decoy compounds. Thus, the deviation of docking scores with multiple proteins should be a measure of drug-likeness for compound affinity.



1. INTRODUCTION

Drug-likeness is a fuzzy idea. Beginning several decades ago, medicinal chemists have understood that drugs have some common features.[1] A rational definition of drug-likeness was proposed by Lipinski, and the set of four criteria, all involving multiples of five, became known as the rule of five.[2] After the rule of five, other definitions of drug- or lead-likeness were proposed.[3,4] One of the popular ideas is the rule of three for fragment-based drug development.[3] Currently, up to 10 million commercially available compounds are filtered by these drug-likeness rules before actual use in a drug development project.

These definitions of drug-likeness are based on chemoinformatics. Namely, only the chemical structures of drugs were analyzed, but the interaction between the proteins and the drugs were ignored. Thus, the rule of five does not define high affinity with the target protein but defines compounds with a good adsorption property. The rule of three is similar to the rule of five. Since the molecular size of the drug is increased by the golden ratio from its seed compound, the rule of five was modified for filtering lead compounds.[5] These studies are based on one-dimensional descriptors. On the other hand, common molecular scaffolds have been analyzed. Mainly, the frequency of ring systems has been analyzed.[6] In addition, nondrug-like substructures, e.g., reactive functional groups, have been listed in textbooks of medicinal chemistry.[7]

Identifying compounds with drug-likeness as conventionally defined does not select the compounds with high affinity. In the current study, we tried to develop a definition of drug-likeness for high affinity. To this purpose, we investigated the interaction between proteins and drugs using a protein−compound docking program. Of course, the target protein itself was not used in the current study.

2. METHOD

A set of active compounds of a target protein and a set of decoy compounds of the target protein was prepared. Protein sets consisting of multiple protein structures were prepared. The average value (μ) and deviation (σ) of docking scores against these multiple proteins were calculated for each compound. The docking scores were given by a protein−compound docking calculation that was performed by the Sievgene/myPresto protein−compound docking program.[8]

The probability of finding active compounds (P_{active}) at a given μ or σ value was calculated using Bayesian analysis. The result of σ was fitted by a sigmoid curve. Namely

$$P^{fit}{}_{active}(\sigma) = \frac{c}{1 + e^{a(\sigma - b)}} \quad (1)$$

where $P^{fit}{}_{active}(\sigma)$, a, b, and c are the probability of finding active compounds at σ and the three constants, respectively. The b value represents the σ value at which the probability of finding active compounds is 50% of the maximum probability [= $P^{fit}{}_{active}(\infty)$].

3. PREPARATION OF MATERIALS

To evaluate our method, we performed a protein−compound docking simulation based on the soluble protein structures registered in the Protein Data Bank (PDB). Two protein sets were prepared. Protein sets A and B consisted of 160 and 53 proteins, respectively. These protein sets did not include the target proteins of the used active compounds. The PDB codes

Received: January 25, 2011
Published: April 27, 2011

dx.doi.org/10.1021/ci200035q | J. Chem. Inf. Model. 2011, 51, 1012–1016

of the proteins in these two subsets are listed in Appendices A and B.

The protein−ligand complex structures were suitable for the docking study, since the ligand pockets were clearly determined. Protein sets A and B were selected from the data sets used in our previous study.[9] Protein set A was prepared as follows: A total of 180 proteins were selected from the PDB; namely, those selected from the database used in the evaluation of the GOLD and FlexX.[10] The protein data set contains a rich variety of proteins and compounds whose structures have all been determined by high-quality experiments with a resolution of less than 2.5 Å. Coordinates of all atoms except hydrogen are supplied, and the atomic structures around the ligand pockets are reliable. From these 180 proteins, the target proteins of the Directory of Useful Decoys (DUD) active compounds were removed. For protein set B, protein set D of the previous study was used.[9] A clustering analysis selected 63 candidate complex structures from protein set A of the current study,[9] and as with set A, the target proteins of the DUD active compounds were removed. Consequently, protein set B consisted of 53 proteins.

For protein sets A and B, the complexes containing a covalent bond between the protein and ligand were removed, since our docking program cannot perform protein−ligand docking when a covalent bond exists between the protein and the ligand. All water molecules and cofactors were removed from the proteins, and all missing hydrogen atoms were added to form the all-atom models of the proteins.

The active compounds and decoy compounds were downloaded from the DUD.[11] There were 40 sets of active compounds and their decoy sets. All compounds were used in the current study. In addition, two random libraries were prepared as decoy sets. One decoy set was a compound library selected from LigandBox compound database.[12] Randomly from LigandBox, 10 000 compounds whose molecular weight is >150 and <350 Da were selected. Another decoy set was the Coelacanth chemical compound library (Coelacanth Corporation, East Windsor, NJ,), which is a random library consisting of 11 050 compounds. These decoy sets were used in our previous studies and are described in detail there.[9]

The atomic charges of the proteins were the same as the atomic charges in AMBER parm99.[13] The docking pocket of each protein was indicated by the coordinates of the original ligand. For flexible docking, the Sievgene/myPresto (http://medals.jp/myPresto/index.html and http://presto.protein.osaka-u.ac.jp/myPresto4/) program was used.[8] The Sievgene program generated up to 100 conformers for each compound. The detail of the docking score is summarized in Appendix C. It takes 3 s to dock one compound against one protein on one core of Xeon 5570 CPU (2.98 GHz).

4. RESULTS AND DISCUSSION

Figures 1, 2, 3, and 4 show the P_{active} profiles for protein sets A and B. Several combinations of protein sets and decoy sets were examined. Results of protein sets A and B with the DUD decoy sets, protein set A with the LigandBox decoy set, and protein set A with the Coelacanth decoy set are shown in Figures 1−4, respectively. Figures 1a, 2a, 3a, and 4a show the P_{active} profiles based on σ values, and Figures 1b, 2b, 3b, and 4b show the P_{active} profiles based on μ values. In Figures 1a−4a, the P_{active} curves show the same trend, and these results show that the number of proteins and choice of proteins does not remarkably change the P_{active} curve.

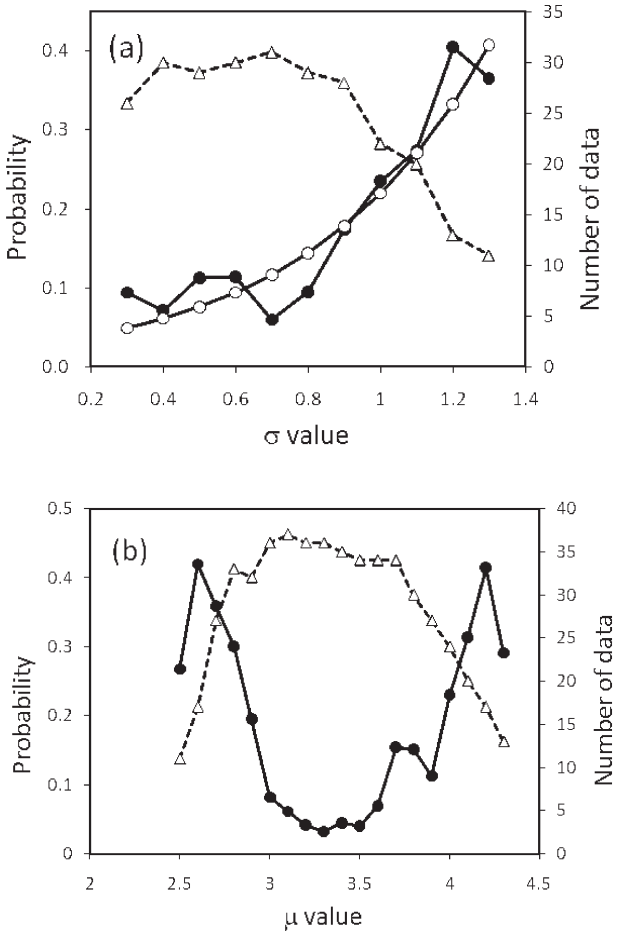


Figure 1. P_{active} profiles obtained by protein set A with the DUD decoy sets. (a) P_{active} based on σ. Filled circles (●), open circles (○), and triangles (△) represent the P_{active} values, the fitting result by eq 1 (P^{fit}_{active}), and the number of data. (b) P_{active} based on μ. Filled circles and triangles represent the P_{active} values and the number of data.

Figures 1a−4a shows that the P_{active} increases by increasing the σ value and that P_{active} can be fitted by eq 1. This result reveals that active compounds show strong affinity with some proteins, even though these proteins are not the true target of the compound, and the same compounds show weak affinity with other proteins. Also, this result shows that inactive compounds show average affinity with many proteins. The optimal fitting parameters of eq 1 were $a = -2.19$, $b = 2.4$, and $c = 4.94$ for protein set A with the DUD decoy set (Figure 1a), $a = -2.72$, $b = 1.29$, and $c = 1.3$ for protein set B with the DUD decoy set (Figure 2a), $a = -2.84$, $b = 2.069$, and $c = 5.3$ for protein set A with the LigandBox decoy set (Figure 3a), and $a = -1.43$, $b = 2.4$, and $c = 1.3$ for protein set A with the Coelacanth decoy set (Figure 4a), respectively. The P_{active} values reached >0.2 at σ = 1.0, 0.8, 1.0, and 1.2 in Figures 1a−4a, respectively. In any case, the P_{active} values reached >0.2 at around $\sigma = 1.0$. As we did not employ target protein structures in the current study, the actual affinity cannot be calculated, but the probability of finding active compounds can be estimated.

This behavior of $P_{active}(\sigma)$ could be explained by the behavior of the universal active probe (UAP). The UAPs are ligands of

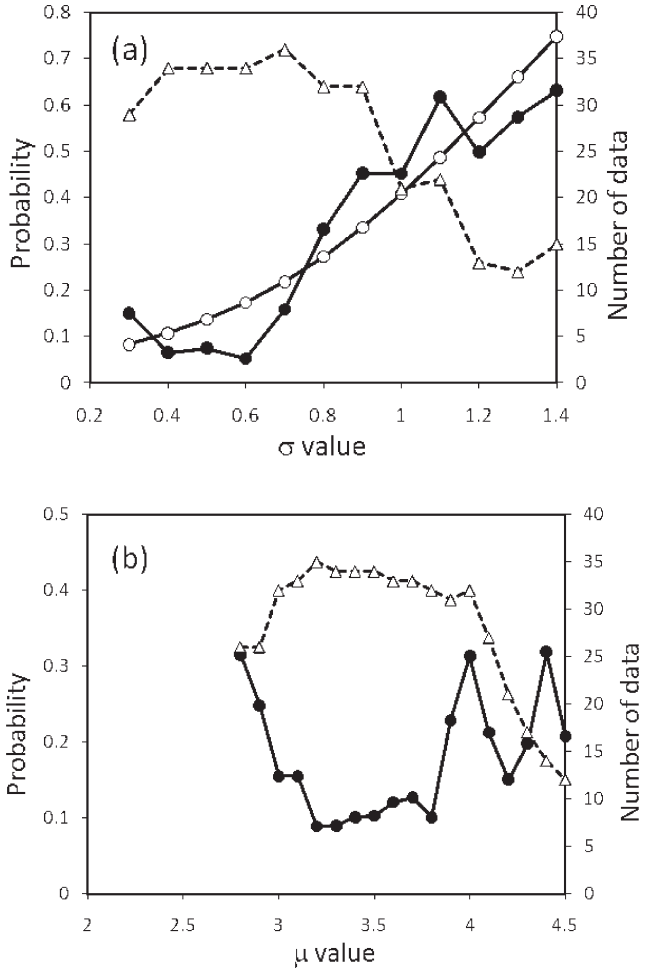


Figure 2. P_{active} profiles obtained by protein set B with the DUD decoy sets. (a) P_{active} based on σ. Filled circles (●), open circles (○), and triangles (△) represent the P_{active} values, the fitting result by eq 1 (P^{fit}_{active}), and the number of data. (b) P_{active} based on μ. Filled circles and triangles represent the P_{active} values and the number of data.

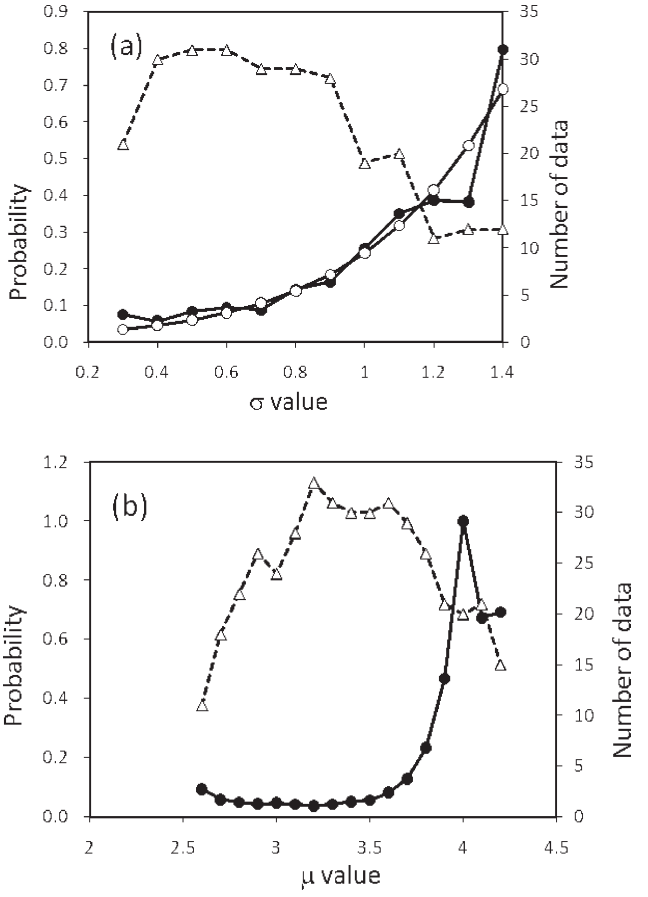


Figure 3. P_{active} profiles obtained by protein set A with the LigandBox decoy set. (a) P_{active} based on σ. Filled circles (●), open circles (○), and triangles (△) represent the P_{active} values, the fitting result by eq 1 (P^{fit}_{active}), and the number of data. (b) P_{active} based on μ. Filled circles and triangles represent the P_{active} values and the number of data.

proteins (known drugs, ligands extracted from the protein–compound complex structures, and active compounds of the DUD that were used in the current study). The UAP is a drug-like compound. In structure-based drug screening, the database enrichment of the true active compounds is proportional to that of the UAPs. The UAP can bind the protein even though the protein is not the true target of the UAP itself. The active compounds in the current study could be the UAPs, and these can bind off-target proteins.

In Figures 1b and 2b, the P_{active} curve has a parabolic shape. This result indicates that some active compounds have a strong affinity with many proteins even though these proteins are not the true target of the compound, and the other active compounds show weak affinity with many proteins. Also, this result indicates that inactive compounds have average affinity with many proteins. On the contrary, the P_{active} curve based on μ is an upward-sloping curve in Figure 3b, and the P_{active} curve based on μ is a downward-sloping curve in Figure 4b. These trends were inconsistent with the parabolic shape of the P_{active} curve shown in Figures 1b and 2b. Thus, it appears the P_{active} curve based on μ is not so meaningful. We do not have a rational explanation for this behavior.

In Figures 1–4, the number of data is the number of targets that have data at a given σ or μ value. The maximum number of data is 40. The error of P_{active} increases when the number of data decreases. The bumps at both ends of P_{active} should not be considered meaningful.

Figure 5 shows the P_{active} profile for protein set A with the standard deviations. The error bars represent the standard deviations of the P_{active} obtained for the DUD 40 sets. The wide error bars suggest that the general trend is upward sloping, but there could be exceptions. We also examined additional four cases. Namely, the P_{active} profiles were calculated for the Coelacanth decoy set with the ligands of protein set A, the Coelacanth decoy set with the ligands of G-protein coupled receptors (GPCRs), the LigandBox decoy set with the ligands of protein set A, and the LigandBox decoy set with the ligands of GPCRs. The used ligands of the GPCRs were total 67 compounds, these are exactly the same as the ligands used in our previous study.[14] Upward-sloping curves of P_{active} profiles were observed for three cases, but a downward-sloping curve was observed for one case that is the Coelacanth decoy set with the ligands of GPCRs. The P_{active} profiles of these four cases are provided as the Supporting Information.

The P_{active} profile could depend on the selectivity of the active compounds. The selectivities of kinase drug and GPCR drug have been well-known. But in the current study, the

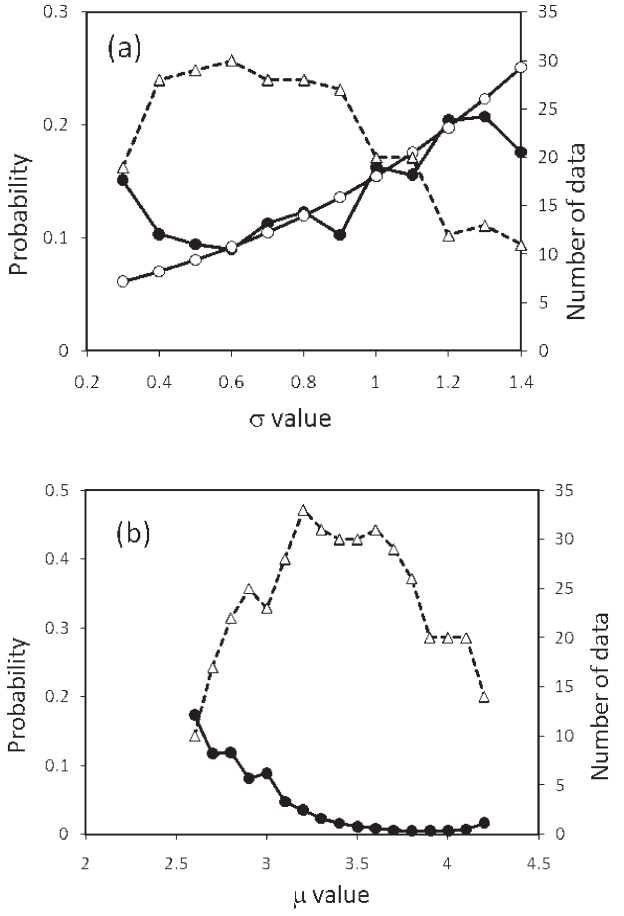


Figure 4. P_{active} profiles obtained by protein set A with the Coelacanth decoy set. (a) P_{active} based on σ. Filled circles (●), open circles (○), and triangles (△) represent the P_{active} values, the fitting result by eq 1 (P^{fit}_{active}), and the number of data. (b) P_{active} based on μ. Filled circles and triangles represent the P_{active} values and the number of data.

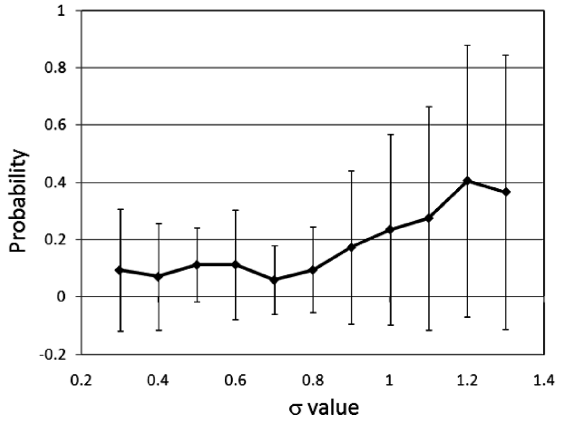


Figure 5. P_{active} based on σ profiles with error bars obtained by protein set A with the DUD decoy sets.

selectivity for many (wide) kinds of proteins is in question. There have been some reports on the selectivity,[15,16] but the number of examined proteins was limited or the number of tested compounds was too small. Thus, this problem should be discussed as a future issue.

5. CONCLUSIONS

We examined the difference of docking score distributions of active compounds and decoy compounds using protein−compound docking against multiple proteins. Each compound was docked to multiple proteins that were not its target proteins, and the set of docking scores was analyzed. The probability of a compound being active (P_{active}) increased as the docking score deviation (σ value) increased, and P_{active} values could be fitted by eq 1. This result shows that active compounds show strong affinity with some proteins even though these proteins are not the true target of the compound, and the same compounds show weak affinity with other proteins. Also, this result shows that inactive compounds show average affinity with many proteins. This trend did not change so much for selection and number of proteins. In any case we examined, the P_{active} values reached >0.2 at around $\sigma = 1.0$. This means that the compounds with $\sigma > 1$ are potentially active compounds.

P_{active} calculated by eq 1 could give the drug-likeness score that suggests its affinity. The protein−compound docking Sievgene program takes 3 s for the docking of 1 compound against one protein. Thus, the P_{active} of a compound could be calculated in 3 or 8 min by 1 core CPU when 53 or 160 proteins were used for docking analysis.

APPENDIX A

The selected 160 proteins (protein set A) were as follows: 1a28, 1ai5, 1b58, 1bqq, 1c83, 1cbx, 1cdg, 1com, 1coy, 1cvu, 1d3h, 1dog, 1epb, 1fen, 1fki, 1fl3, 1gc7, 1hfc, 1hos, 1jap, 1lcp, 1ldm, 1mbi, 1mdr, 1mld, 1mmq, 1mrg, 1mup, 1ngp, 1okl, 1pbd, 1pdz, 1pso, 1qbu, 1qpq, 1tng, 1xie, 1yee, 2ack, 2ada, 2cmd, 2ctc, 2fox, 2ifb, 2pk4, 3cpa, 3tpi, 4aah, 4lbd, 12as, 16gs, 18gs, 1a42, 1a4g, 1a4q, 1abe, 1abf, 1aco, 1ady, 1aer, 1aoe, 1apt, 1apu, 1aqw, 1asz, 1atl, 1aux, 1b76, 1b9v, 1bdg, 1bma, 1byb, 1byg, 1c1e, 1c5c, 1cbs, 1ckp, 1cps, 1csn, 1d0l, 1dd7, 1dg5, 1dhf, 1dr1, 1ebg, 1eed, 1efv, 1ejn, 1epo, 1ets, 1f0r, 1f0s, 1f3d, 1fkg, 1glg, 1glp, 1gol, 1gtr, 1hck, 1hdc, 1hpx, 1hsb, 1hsl, 1hyt, 1ivb, 1l3f, 1lah, 1lic, 1lna, 1lst, 1mmu, 1mts, 1nco, 1nis, 1nks, 1phd, 1phg, 1poc, 1ppc, 1pph, 1pyg, 1qbr, 1qh7, 1rds, 1rne, 1rnt, 1rob, 1ses, 1snc, 1so0, 1srj, 1tlp, 1tmn, 1tnh, 1tni, 1tnl, 1tyl, 1xid, 2aac, 2aad, 2cht, 2cpp, 2gbp, 2gbp, 2gss, 2qwk, 2tmd, 2tmn, 3cla, 3hvp, 3pgh, 3pgt, 3r1r, 4est, 4phv, 5abp, 5cpp, 5er1, 6rnt, and 7tim.

APPENDIX B

The selected 53 proteins (protein set B) were as follows: 18gs, 2gss, 1hpx, 2tmn, 1a28, 1ai5, 1b58, 1bqq, 1c83, 1cbx, 1cdg, 1com, 1coy, 1cvu, 1d3h, 1dog, 1epb, 1fen, 1fki, 1fl3, 1hfc, 1hos, 1jap, 1lcp, 1ldm, 1mbi, 1mdr, 1gc7, 1mld, 1mmq, 1mrg, 1mup, 1ngp, 1okl, 1pbd, 1pdz, 1pso, 1qbu, 1qpq, 1tng, 1xie, 1yee, 2ack, 2ada, 2cmd, 2ctc, 2fox, 2ifb, 2pk4, 3cpa, 3tpi, 4aah, and 4lbd.

APPENDIX C

The sievgene score is determined as

$$S_{\Delta G} = c_{rot} \cdot N_{rot} + c_{AV} \cdot (E_{ASA} + E_{vdW}) + c_{ele} \cdot E_{ele} + c_{hyd} \cdot E_{hyd} + c_{intra\text{-}vdW} \cdot E_{intra\text{-}vdW}$$

where N_{rot}, E_{ASA}, E_{vdW}, E_{ele}, E_{hyd}, and $E_{intra\text{-}vdW}$ represent the number of rotatable bonds of docked compound, the hydrophobic energy due to the accessible surface area, the vdW energy, the protein−ligand Coulomb potential, the hydrogen-bond

dx.doi.org/10.1021/ci200035q |J. Chem. Inf. Model. 2011, 51, 1012–1016

energy, and the intramolecule vdW energy of the ligand. Also, c_{rot}, c_{AV}, c_{ele}, c_{hyd}, and $c_{intra\text{-}vdW}$ are the optimized coefficients for each energy term. For each atom type, the sum of E_{ASA} and E_{vdW} gives one grid potential. Sievgene utilizes the grid potential to calculate each energy term except the intramolecule interaction. In this study, the mesh size of 60 × 60 × 60 was adopted.

■ ASSOCIATED CONTENT

Supporting Information. The P_{active} profiles were calculated for the Coelacanth decoy set with the ligands of protein set A, the Coelacanth decoy set with the ligands of GPCRs, the LigandBox decoy set with the ligands of protein set A, and the LigandBox decoy set with the ligands of GPCRs. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: y-fukunishi@aist.go.jp. Telephone: +81-3-3599-8290.

■ ACKNOWLEDGMENT

This work was supported by grants from the New Energy and Industrial Technology Development Organization of Japan (NEDO) and the Ministry of Economy, Trade, and Industry (METI) of Japan.

■ REFERENCES

(1) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. *J. Combin. Chem.* **1999**, *1*, 55–68.

(2) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235–249.

(3) 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.

(4) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'Rule of Three' for fragment-based lead discovery?. *Drug Discovery Today* **2003**, *8*, 876–877.

(5) Orita, M.; Ohno, K.; Niimi, T. Two "Golden ratio" indices in fragment-based drug discovery. *Drug Discovery Today* **2009**, *14*, 321–328.

(6) Fujii, I.; Sugaya, N.; Nakano, T.; Hasegawa, S.; Yamamoto, M.; Kaminuma, T.; Hirayama, N. Essential chemical characteristics for drugs. *Chem-Bio Inf. J.* **2001**, *1*, 18–22.

(7) Wermuth, C. G. *The practice of medicinal chemistry*; Academic Press: Burlington, MA, 2008.

(8) Fukunishi, Y.; Mikami, Y.; Nakamura, H. Similarities among receptor pockets and among compounds: Analysis and application to in silico ligand screening. *J. Mol. Graphics Modell.* **2005**, *24*, 34–45.

(9) Fukunishi, Y.; Kubota, S.; Nakamura, H. Noise reduction method for molecular interaction energy: application to in silico drug screening and in silico target protein screening. *J. Chem. Inf. Model.* **2006**, *46*, 2071–2084.

(10) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein-ligand interaction. *Proteins* **2002**, *49*, 457–471.

(11) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(12) Fukunishi, Y.; Sugihara, Y.; Mikami, Y.; Sakai, K.; Kusudo, H.; Nakamura, H. Advanced in-silico drug screeing to achieve high hit ratio-development of 3D-compound database. *Synthesiology* **2009**, *2*, 60–68.

(13) Case, D. A.; Darden, T. A.; Cheatham, T. E., III.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER 8*, University of California, San Francisco: San Francisco, CA, 2004.

(14) Fukunishi, Y; Hojo, S.; Nakamura, H. An efficient in silico screening method based on the protein-compound affinity matrix and its application to the design of a focused library for cytochrome P450 (CYP) ligands. *J. Chem. Inf. Comput. Sci.* **2006**, *46*, 2610–22.

(15) Steindl, T. M.; Schuster, D.; Laggner, C.; Langer, T. Parallel screening: a novel concept in pharmacophore modeling and virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 2146–2157.

(16) Tobita, M.; Horiuchi, K.; Araki, K.; Nemoto, M.; Shimada, H.; Nishikawa, T. BirdsAnts: a protein-small molecule interaction viewer. *Chem-Bio Inf. J.* **2006**, *6*, 17–28.