

## POT-DMC: A Virtual Screening Method for the Identification of Potent Hits

Jeffrey W. Godden,<sup>†</sup> Florence L. Stahura,<sup>†</sup> and Jürgen Bajorath<sup>†,‡,\*</sup>

Albany Molecular Research - Bothell Research Center (AMRI-BRC), 18804 North Creek Parkway, Bothell, Washington 98011, and Department of Biological Structure, University of Washington, Seattle, Washington 98195

Received June 22, 2004

**Abstract:** A method for ligand-based virtual screening (LBVS), dynamic mapping of consensus positions (DMC), has been extended to take different potency levels of template compounds into account. This potency scaling technique is designed to tune search calculations toward the detection of increasingly potent hits. LBVS analysis of three different compound classes confirmed the ability of potency-scaled DMC (POT-DMC) to identify active database compounds with higher potency than conventional calculations.

In pharmaceutical research, virtual screening of large compound databases has become an important hit identification tool.<sup>1</sup> For LBVS, which utilizes known active compounds as search templates, a variety of molecular similarity-based methods<sup>2</sup> have been developed or adapted including, among others, molecular fingerprints,<sup>3,4</sup> clustering techniques,<sup>5,6</sup> and partitioning methods.<sup>6,7</sup> A major goal of LBVS is the identification of molecules having core structures different from known leads or drugs but similar activity.<sup>1</sup> This represents an important difference from QSAR methods where congeneric molecules are usually analyzed in order to identify or design analogues with increased potency.<sup>8</sup> However, in LBVS algorithms, relative compound potency has thus far rarely been considered as a search parameter, although QSAR models have also been adapted for virtual screening.<sup>8–10</sup> We have begun to address this issue by extending the recently developed DMC approach,<sup>11</sup> which combines elements of partitioning algorithms and bit string methods, to include compound potency as a new parameter during similarity analysis. This is done in order to increase the probability of finding the most potent hits in screening databases.

The concept of DMC and POT-DMC is illustrated in Figure 1a. In contrast to popular cell-based partitioning methods that operate in low-dimensional descriptor reference spaces,<sup>7</sup> DMC utilizes dimension extension of simplified descriptor spaces to distinguish sets of active molecules from other database compounds. This separation is facilitated by finding consensus positions for active compounds in descriptor spaces of increasing dimensionality. During dimension extension, the number of database compounds that closely map to active templates decreases and only similar compounds are retained. Simplified reference spaces are generated by

binary descriptor transformation.<sup>11</sup> This process converts descriptors with continuous value ranges into a binary format based on the statistical medians of their value distributions in the screening database. Each test compound is assigned a “1” for a descriptor if its value is larger or equal to the median or a “0” if it is smaller. This binary model makes it possible to generate descriptor bit strings for mapping of compounds. A consensus position is defined by a descriptor vector in chemical space composed of those bit settings that are identical for all template compounds. Dimension extension is achieved by establishing consensus positions that no longer require identical descriptor settings for all templates. As reported herein, dimension extension levels 1, 2, and 3 allow 10%, 20%, and 30% variability in descriptor bits settings, respectively. For example, if 10 template compounds are available, a descriptor is accepted at dimension extension level 1, if 9 of 10 compounds have the same binary setting (either 0 or 1). This process increases the number of descriptors in consensus bit strings and hence the resolution of the reference space. Ultimately, molecules that map to activity class-specific positions after elimination of most of the database compounds are considered potential hits. In benchmark calculations on several drug classes, DMC was found to produce significant rates of up to 74%.<sup>11</sup>

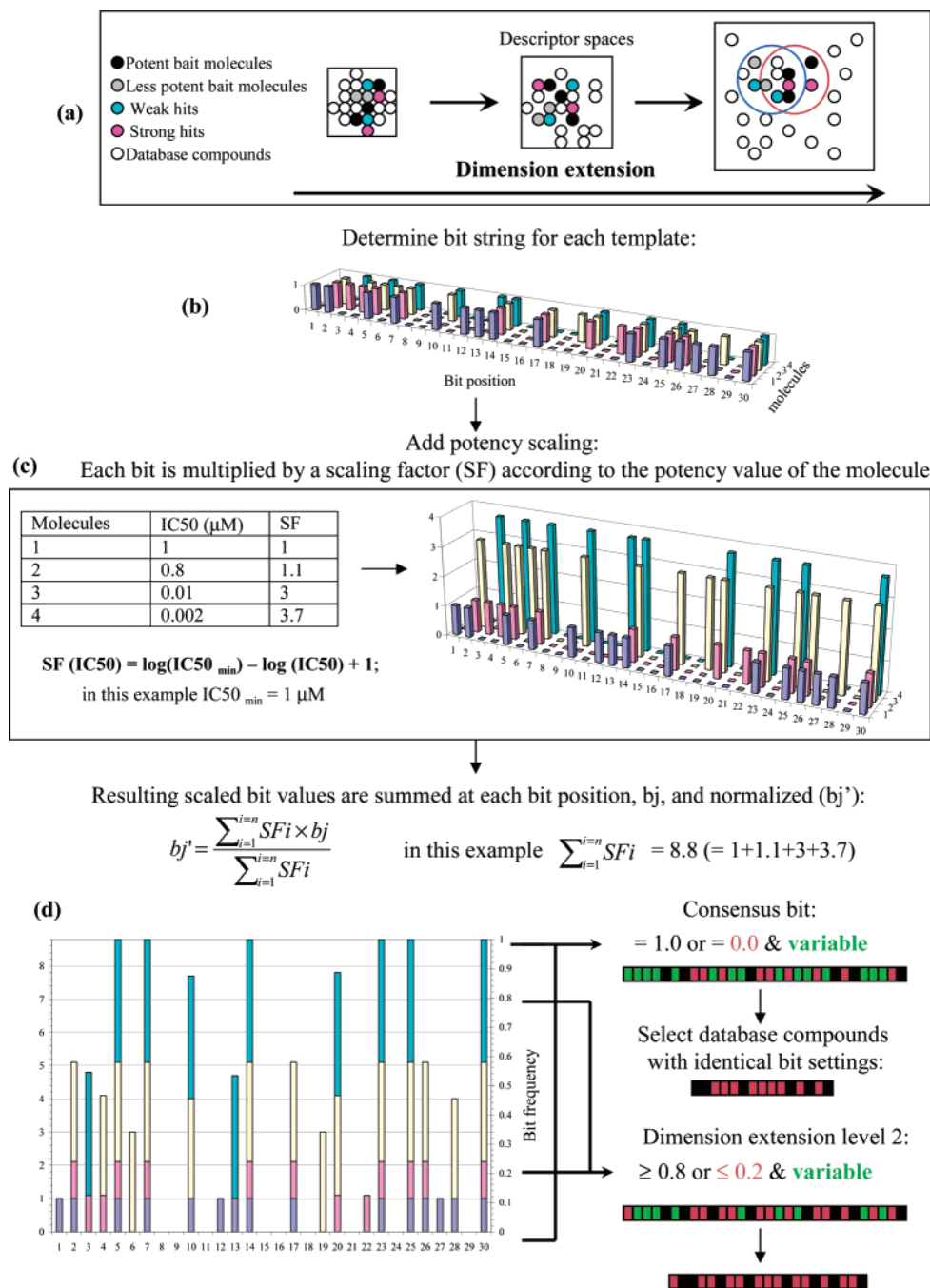
As illustrated in Figure 1a, the principal idea behind POT-DMC is the generation of consensus positions that assign increasing weight to high-potency templates and are more likely to be matched by most potent database hits. Importantly, however, contributions from less potent templates are also considered, which ensures that structural information of the entire template set is taken into account during LBVS. Compound potency is added as a search parameter through a scaling technique. Thus, once descriptor bit strings are determined for each template (Figure 1b), they are scaled according to template potency, as shown in Figure 1c. For this purpose, logarithmic scaling factors (SFs) are calculated (so that SF for the weakest potency is 1). The use of logarithmic SFs ensures linear scaling over the entire potency range and avoids that the calculations are completely dominated by the most potent compounds. Scaled bit values are summed and normalized to obtain potency-scaled descriptor frequencies (Figure 1d) that are then used to calculate consensus positions. The initial consensus position (level 0) is not affected by scaling, which becomes effective during dimension extension when POT-DMC produces consensus positions that are increasingly influenced by contributions from the most potent templates (and thus differ from DMC consensus positions).

The POT-DMC method has been tested and compared to standard DMC on three different compound classes including CCR5 chemokine receptor antagonists (CCR5),<sup>12</sup> serotonin receptor agonists (5-HT<sub>3</sub>),<sup>13</sup> and gonadotropin releasing hormone agonists (GnRH; assembled in-house from the patent literature). Each of these classes contains structurally diverse compound series covering a wide potency range. Each class was randomly divided into two similarly sized sets spanning

\* To whom any correspondence should be addressed at AMRI-BRC: phone: (425) 424-7297, fax: (425) 424-7299, e-mail: jurgen.bajorath@albmo.com.

<sup>†</sup> Albany Molecular Research - Bothell Research Center.

<sup>‡</sup> University of Washington.

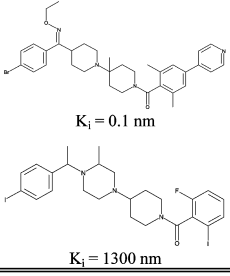
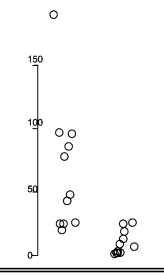
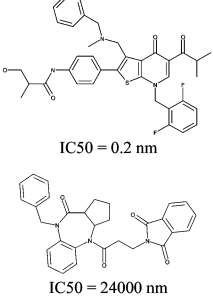
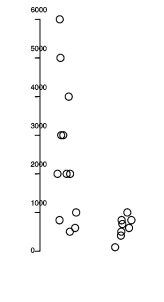
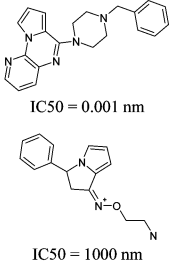
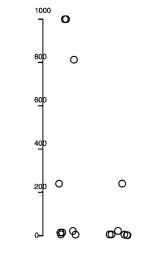


**Figure 1.** POT-DMC. The figure summarizes the POT-DMC approach, as described in the text. I(a) The blue and red circles outline compound selections based on DMC and POT-DMC calculations, respectively. (b) Bit strings for 30 binary-transformed descriptors (set to 0 or 1) and four test molecules (each in a different color). (c) The potency scaling approach is illustrated for these four templates with hypothetical potency.  $IC_{50_{min}}$  refers to the compound having lowest potency. (d) The bit frequency profile for the template set (color-coded by compound) and the initial consensus bit string (black/red, = 1.0 or = 0.0, i.e., permitting no bit variability) and the consensus bit string for the second dimension extension level ( $\geq 0.8$  or  $\leq 0.2$ , 20% bit variability allowed for descriptors set to either 1 or 0).

the entire potency range, and each of these sets was used once as the templates and potential hits (for both DMC and POT-DMC) in order to reduce the probability of chance effects. Potential hits were added to a large compound collection from various medicinal chemistry vendors containing 1.34 million molecules.<sup>14</sup> As the descriptor pool, a set of ~100 previously reported 1D, 2D, and implicit 3D molecular descriptors<sup>14</sup> was used and binary-transformed based on their database medians. Results of our virtual screening calculations are reported in Table 1.

In both DMC and POT-DMC calculations, the number of database compounds that copartitioned with correctly identified hits decreased sharply during dimension extension to produce satisfactory to significant hit rates (e.g., 80–100% for CCR5) beginning at extension levels 2 (CCR5) or 3 (GnRH, 5-HT3). During dimension extension, the number of descriptors defining consensus positions was similar for DMC and POT-DMC for each class but varied between classes. While DMC and POT-DMC hit rates were overall also similar, the potency distribution of correctly identified hits revealed some

**Table 1.** Virtual Screening Trials<sup>a</sup>

Activity class (# compounds)	Example structures (most and least potent)			Extension level 0	Extension level 1 (hit rate in %)	Extension level 2 (hit rate in %)	Extension level 3 (hit rate in %)	Potency distribution of hits	
								DMC	POT DMC
CCR5 (98)		DMC	#dbc	23442	59	2	0		
			#hits	48	27 (31)	12 (86)	6 (100)		
			<Pot> K <sub>i</sub> (nm)	46.5	37.8	70.3	9.8		
		POT-DMC	#dbc	23442	44	2	0		
			#hits	48	23 (34)	10 (83)	5 (100)		
			<Pot> K <sub>i</sub> (nm)	46.5	37.5	11.8	7.3		
GnRH (100)		DMC	#dbc	890290	63819	484	98		
			#hits	47	29	11 (2.2)	7 (6.7)		
			<Pot> IC <sub>50</sub> (nm)	3060	3120	2015	1055		
		POT-DMC	#dbc	890290	189690	10820	49		
			#hits	47	37	9 (0.08)	5 (9.3)		
			<Pot> IC <sub>50</sub> (nm)	3060	2245	636	260		
5-HT3 (38)		DMC	#dbc	2858	923	355	28		
			#hits	18	14	9 (2.5)	3 (9.7)		
			<Pot> IC <sub>50</sub> (nm)	272.5	348.5	293.0	11.3		
		POT-DMC	#dbc	2858	643	91	4		
			#hits	18	11	6 (6.2)	1 (20)		
			<Pot> IC <sub>50</sub> (nm)	272.5	63.8	91.9	0.004		

<sup>a</sup> #hits reports the number of hits that mapped to template consensus positions and were thus correctly identified, and #dbc the number of other detected database compounds (considered false-positives). <Pot> gives the average potency of correctly identified hits that was calculated as a statistical measure of the potency distribution of distinct hit populations identified in the presence and absence of scaling. For each class, the average of two virtual screening trials is reported, and the potency distribution is graphically presented for the POT-DMC run yielding highest average potency at extension level 2. Hit rates were calculated when the total number of detected compounds was smaller than 100 (which would represent a reasonably sized selection set for many LBVS applications).

significant differences. POT-DMC showed a clear trend to increase the average potency of identified hits during dimension extension. Across activity classes and extension levels, average hit potency was consistently higher for POT-DMC than DMC, sometimes by up to an order of magnitude (which is significant for an average). The graphical analysis of potency distributions in Table 1 confirmed that POT-DMC recognized more potent hits than DMC and did not detect weak ones.

On the basis of these findings, we conclude that relative compound potency has been successfully integrated as a search criterion into an LBVS method to identify potent hits. A key aspect of the POT-DMC approach presented herein is the generation of activity class- and potency-dependent consensus positions in descriptor space for mapping of active compounds. The logarithmic scaling technique designed to incorporate relative compound potency as a search parameter should also be applicable to other partitioning and LBVS schemes.

**Acknowledgment.** We wish to thank Asim K. Debnath, Lindsley F. Kimball Research Institute, New York Blood Center, for making the CCR5 compound dataset available to us.

## References

- Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.
- Johnson, M. A.; Maggiora, G. M.; Eds. *Concepts and applications of molecular similarity*; Wiley: New York, 2000.
- Xue, L.; Godden, J. W.; Bajorath, J. Mini-fingerprints: design principles and generation of novel prototypes based on information theory. *SAR QSAR Environ. Res.* **2003**, *14*, 27–40.
- Mason, J. S.; Cheney, D. L. Library design and virtual screening using multiple 4-point pharmacophore fingerprints. *Pac. Symp. Biocomput.* **2000**, *5*, 576–587.
- Willett, P.; Wintermann, V.; Bawden, D. Implementation of nonhierarchical cluster analysis methods in chemical information systems: selection of compounds for biological testing and clustering of substructure search output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109–118.
- Stahura, F. L.; Bajorath, J. Partitioning methods for the identification of active molecules. *Curr. Med. Chem.* **2003**, *8*, 707–715.
- Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspect. Drug Discovery Des.* **1998**, *9*, 339–353.
- Hopfinger, A. J.; Reaka, A.; Venkatarangan, P.; Duca, J. S.; Wang, S. Construction of a virtual high throughput screen by 4D-QSAR analysis: application to a combinatorial library of glucose inhibitors of glycogen phosphorylase b. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1151–1160.
- Tropsha, A.; Zheng, W. Identification of the descriptor pharmacophores using variable selection QSAR: applications to database mining. *Curr. Pharm. Des.* **2001**, *7*, 599–612.

- (10) Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J. P.; Kohn, H.; Tropsha, A. Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. *J. Med. Chem.* **2004**, *47*, 2356–2364.
- (11) Godden, J. W.; Furr, J. R.; Xue, L.; Stahura, F. L.; Bajorath, J. Molecular similarity analysis and virtual screening by mapping of consensus positions in binary-transformed chemical descriptor spaces with variable dimensionality. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 21–29.
- (12) Debnath, A. K. Generation of predictive pharmacophore models for CCR5 antagonists: study with piperidine- and piperazine-based compounds as a new class of HIV-1 entry inhibitors. *J. Med. Chem.* **2003**, *46*, 4501–4515.
- (13) Daveu, C.; Bureau, R.; Baglin, I.; Prunier, H.; Lancelot, J.-C.; Rault, S. Definition of a pharmacophore for partial agonists of serotonin 5-HT<sub>3</sub> receptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 362–369.
- (14) Godden, J. W.; Xue, L.; Kitchen, D. B.; Stahura, F. L.; Scherm-erhorn, E. J.; Bajorath, J. Median partitioning: a novel method for the selection of representative subsets from large compound pools. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 885–893.

JM049505G