# Exploring Target-Selectivity Patterns of Molecular Scaffolds

Ye Hu and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

ABSTRACT We investigate the question of whether target-selective molecular scaffolds can be identified on the basis of currently available compound activity data. Starting from a pool of 17745 public domain compounds with activity annotations for 433 human targets, we ultimately identify, through a selectivity classification and database-mining approach, 42 molecular scaffolds represented by multiple compounds that are highly selective for a particular target over one or more others. In many other cases, individual compounds representing unique scaffolds are target-selective. Hence, currently available public domain compound selectivity data are sparse. However, we also identify selectivity patterns that evolve around specific targets and are formed by multiple target-selective scaffolds. These scaffolds should provide interesting starting points for further chemical exploration.

KEYWORDS Molecular selectivity, privileged substructures, target family selective molecular scaffolds, target-selective scaffolds, selectivity patterns, compound database mining
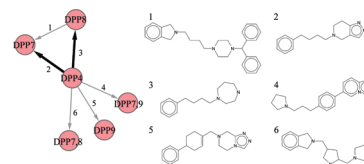
In medicinal chemistry, "privileged substructures",[1] that is, chemotypes that bind with high preference to a family of targets, have been—and continue to be—intensely studied. In many instances, substructures considered to be target class-selective on the basis of frequency of occurrence analysis have also been detected in compounds active against other target families;[2] hence, the existence of truly privileged structural motifs has been controversial.[2]

Recently, we have carried out a large-scale analysis of public domain compound data to investigate whether target class-selective molecular scaffolds exist.[3] To avoid potential caveats of occurrence frequency-based analysis, we searched for compounds with multiple activity annotations and formed pairs of biological targets that were "connected" by at least five active compounds. This target pair information was then organized in a compound-based target network that enabled the identification of different target communities. From these compounds, conventional hierarchical scaffolds[4] were isolated, and scaffolds were determined that exclusively occurred in one of the target communities formed by the network. The approach is summarized on the left side in Figure 1.

For this target pair-based analysis, BindingDB[5] was found to be a comprehensive public domain source of bioactivity data. For example, by systematically analyzing currently available PubChem[6] confirmatory bioassays, only three target pairs were identified that met the selection criterion. Of 17745 compounds available in BindingDB with activity annotations against a total of 433 human targets, 6343 compounds active against 259 targets met our target pair selection criterion (i.e., five or more shared ligands),

yielding a total 520 target pairs organized into 18 target communities. From these 6343 compounds, a total of 206 target community-selective scaffolds were identified, that is, scaffolds that only occurred in one of 18 communities (Figure 1). We also calculated a pairwise potency-based selectivity ratio for compounds representing these scaffolds, which indicated that a subset of these scaffolds had the potential to yield selective compounds, at least at the level of target pairs.[3]

In light of these findings, a logical follow-up question thus became if there might also be truly target-selective scaffolds present among community-selective ones. Target-selective scaffolds, that is, scaffolds that exclusively yield target-selective compounds, would be of high interest for medicinal chemistry research. Hence, we have investigated this question and report the results herein.

To make the analysis of target-selective scaffolds as comprehensive as possible, we decided to revise the previous target pair and scaffold selection approach, as illustrated on the right side in Figure 1. Therefore, we applied a more stringent target pair selection criterion by requiring not only at least five shared ligands but also at least five scaffolds representing shared ligands. A total of 220 human targets yielding 428 target pairs met these requirements and are reported in Table S1 of the Supporting Information with their BindingDB target IDs. This target pair information was
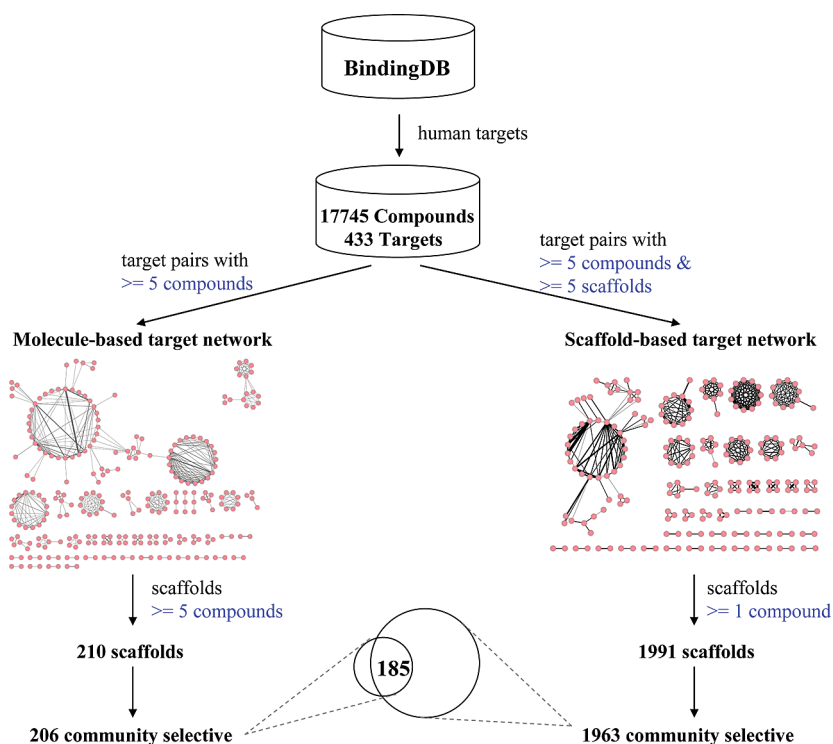
Figure 1. Target communities and community-selective scaffolds. Shown is an overview of alternative approaches to establish target communities on the basis of compound activity data and isolate community-selective scaffolds, which provide a basis for the identification of target-selective scaffolds.

then organized in a scaffold-based target network, as illustrated on the right in Figure 1. In this network, targets are connected if compounds active against them represent at least five scaffolds. We found that this scaffold-based target network further refined the formation of target communities as compared to the previous compound-based target network. In the scaffold-based network, 21 well-defined communities containing at least three targets were found (rather than 18). The target, compound, and scaffold composition of these 21 communities are reported in Table 1, and the scaffold-based network with target community annotations is shown in Figure S1 of the Supporting Information. After a more stringent target pair criterion was applied, we then relaxed the scaffold selection criterion by accepting any scaffold (and not only scaffolds represented by at least five compounds), which yielded a total of 1991 scaffolds, 1963 of which occurred in only one of 21 communities. These community-selective scaffolds were active against 174 targets in 405 target pairs and also included 185 of the 206 community-selective scaffolds previously identified from the compound-based target network (Figure 1, left side). The remaining 21 scaffolds occurred in more than one of the 21 communities in the scaffold-based network.

The 1963 community-selective scaffolds were then ranked on the basis of the median absolute selectivity ratio (|pSR|) of compounds that they represent for established target pairs. The absolute selectivity ratio of a compound for a target pair is simply given by the positive difference of its logarithmic potency values against the two targets. Accordingly,

median |pSR| values ≥ 1 and ≥ 2 indicate that at least half of the compounds represented by a scaffold have at least a 10- and 100-fold potency difference for one target over another, respectively. Figure S2 of the Supporting Information shows the distribution of scaffolds over median |pSR| values, the number of compounds that they represent, and the target pairs that these compounds are active against. Of the 1963 community-selective scaffolds, 1026 scaffolds had a median |pSR| ≥ 1, and 329 scaffolds had a median |pSR| ≥ 2. Thus, a significant number of scaffolds corresponded to highly selective compounds. However, 1350 scaffolds were found to represent a single compound, 1049 scaffolds were found to be active against a single target pair, and 785 scaffolds were found to correspond to both a single molecule and a target pair. Thus, this distribution reflects a notable degree of data incompleteness, which generally affects the systematic analysis of target−ligand interactions.[7] Hence, when more compounds representing individual scaffolds and more measurements become available, the number of selective scaffolds is expected to decrease. However, among the 329 highly selective scaffolds with median |pSR| ≥ 2, there were also 50 scaffolds that represented multiple compounds active against multiple target pairs (Figure S2 of the Supporting Information), which represented particularly interesting scaffolds for further analysis.

Community-selective scaffolds were further classified according to different selectivity threshold levels of the compounds that they represent, that is, at least 10-, 50-, or 100-fold selectivity. The classification scheme is illustrated in

Table 1. Composition of Target Communities[a]

		no. of			
community	target family	targets	target pairs	compounds	scaffolds
1	tyrosine kinases and cytochrome P450 enzymes	50	100	2128	782
2	serine proteinases	12	34	545	229
3	protein kinase C	8	22	72	34
4	carbonic anhydrases	11	55	327	87
5	phosphodiesterases	11	39	117	47
6	matrix metalloproteinases	10	24	187	56
7	protein kinase B and serine protein kinases	6	11	109	78
8	caspases	9	31	114	49
9	histone deacetylases	8	22	121	68
10	purinergic receptors	6	7	107	54
11	phosphoinositide 3-kinases (PI3Ks)	6	10	46	26
12	GABAA receptors	5	9	8	7
13	opioid receptors	4	6	84	27
14	cathepsins	4	6	307	152
15	dipeptidyl peptidases	4	6	287	105
16	esterases	4	6	238	110
17	polo-like kinases	4	5	35	21
18	sphingosine 1-phosphate (S1P) receptors	3	3	20	9
19	peroxisome proliferator-activated receptors	3	3	61	16
20	steroid receptors	3	3	35	9
21	β-secretases and cathepsin D	3	3	127	66

[a] Target communities extracted from the scaffold-based target network are characterized by the number and nature of the targets and, in addition, by the number of compounds active against pairs of targets and the corresponding scaffolds.
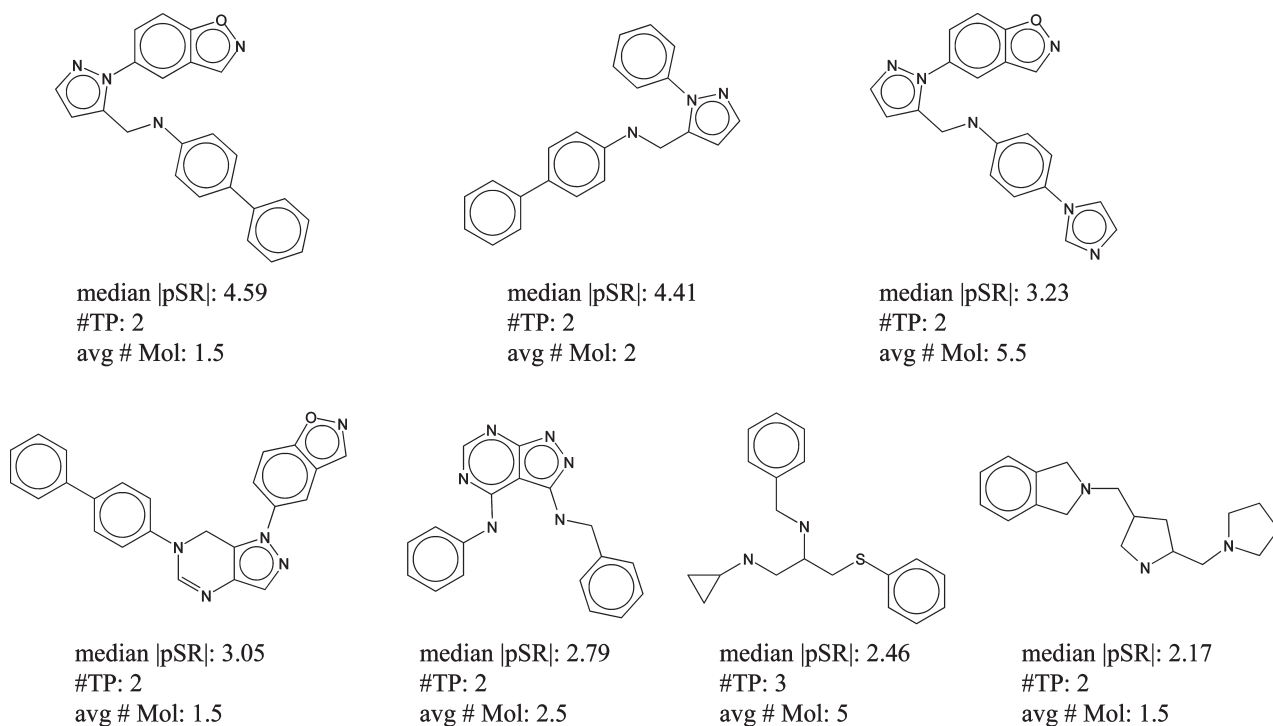

median |pSR|: 4.59
#TP: 2
avg # Mol: 1.5

median |pSR|: 4.41
#TP: 2
avg # Mol: 2

median |pSR|: 3.23
#TP: 2
avg # Mol: 5.5

median |pSR|: 3.05
#TP: 2
avg # Mol: 1.5

median |pSR|: 2.79
#TP: 2
avg # Mol: 2.5

median |pSR|: 2.46
#TP: 3
avg # Mol: 5

median |pSR|: 2.17
#TP: 2
avg # Mol: 1.5

Figure 2. Scaffolds contained in highly selective compounds. Seven scaffolds are shown for which corresponding compounds had median |pSR| ≥ 2 and for which each compound was 50-fold selective for a target over another. For each scaffold, the median median |pSR| value is reported as well as the number of target pairs in which it occurs and the average number of molecules per pair.

Table 2. Target-Selective Scaffolds[a]

scaffolds		no. of		
		scaffolds	targets	target pairs
community-selective		1963	174	405
target-selective	10-fold	**472**	83	66
	50-fold	**250**	65	43
	100-fold	**191**	58	38

[a] Reported is the number of target-selective scaffolds (bold) represented by one or more compounds at different selectivity levels. In addition, the corresponding numbers of targets and target pairs for all community- and target-selective scaffolds are also reported.

Figure S3 of the Supporting Information, and further details are provided in the Methods of the Supporting Information. If a scaffold was always selective for a target over one or more others (in different pairs), it was termed "purely" selective (i.e., a scaffold can be purely selective for more than one target). For the 10-, 50-, and 100-fold selectivity levels, a total of 499, 252, and 191 purely selective scaffolds were identified, respectively. These scaffold sets were compared to the 50 scaffolds with median |pSR| ≥ 2 that represent multiple compounds active against multiple target pairs (Figure S4 of the Supporting Information), revealing an overlap of 11 (10-fold), 7 (50-fold), and 3 (100-fold) scaffolds, respectively. Figure 2 shows the seven purely selective scaffolds for the 50-fold selectivity level. These scaffolds and the corresponding compounds are provided in Table S2 of the Supporting Information.

Having found that community-selective scaffolds had rather different distributions and selectivity profiles, we searched for target-selective scaffolds among the purely selective ones. We considered a scaffold target-selective if it was selective for an individual target over one or more others. Because complex pairwise selectivity relationships can exist for scaffolds in multiple target pairs, the identification of target-selective scaffolds can be complicated. Hence, it was facilitated through a directed graph type method illustrated in Figure S5 of the Supporting Information. Details are provided in Methods of the Supporting Information. In Table 2, the number of target-selective scaffolds is reported. For the 10-, 50-, and 100-fold selectivity levels, 472, 250, and 191 target-selective scaffolds were identified. Hence, most purely selective scaffolds were also target-selective scaffolds. For the 100-fold selectivity level, 149 of 191 target-selective scaffolds only corresponded to a single compound selective for one target over one or two others. The remaining 42 scaffolds were represented by 2−21 compounds and were selective for an individual target over one or two others. These scaffolds are displayed in Figure S6 of the Supporting Information and their target annotations are provided.

Going beyond target selectivity of individual scaffolds, we also asked the question of which target relationships, or selectivity patterns, might be formed by target-selective scaffolds. Therefore, we analyzed the three sets of selective scaffolds reported in Table 2. For the 10-, 50-, and 100-fold selectivity levels, 28 (50), 18 (31), and 19 (23) well-defined target relationships were formed by single (multiple) scaffolds, respectively. As shown for the 50-fold selectivity level
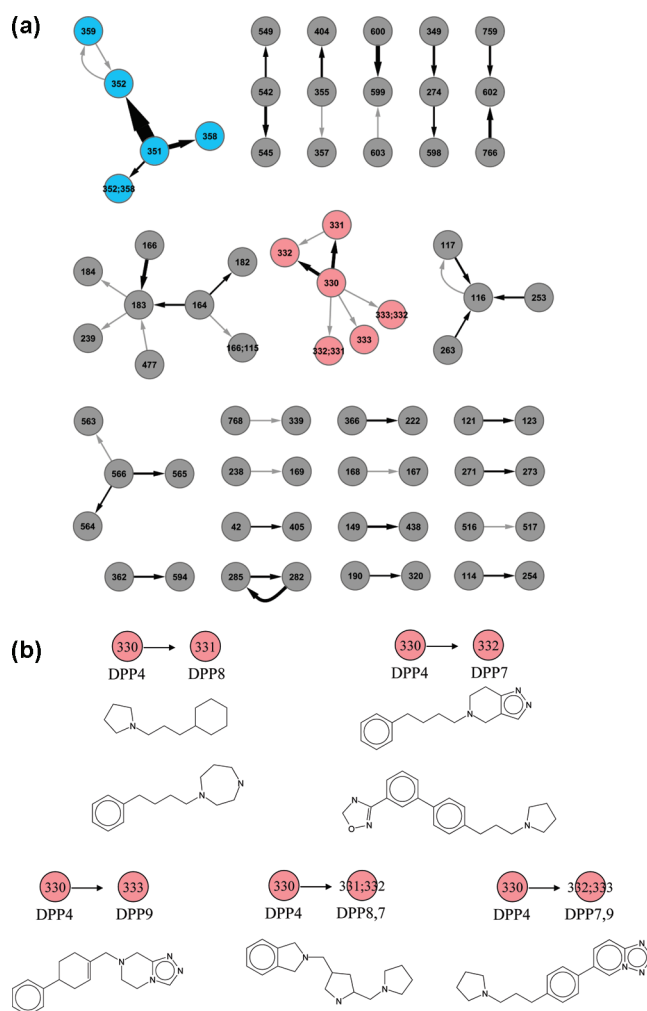

Figure 3. Selectivity patterns. (a) The directed target network for the 50-fold selectivity level is shown displaying different scaffold-based target relationships. The width of directed edges is scaled according to scaffold numbers. When a relationship is formed by a single scaffold, the edge is shown in gray. (b) Scaffolds are shown that yield compounds selective for DDP4 over other DDPs, corresponding to the target cluster with pink nodes in panel a. The two relationships at the top are formed by 10 and 11 scaffolds, respectively, and two representative scaffolds are shown in each case. The three selectivity relationships at the bottom each involve a single scaffold.

in Figure 3a, these relationships can be viewed in a directed target network where nodes (targets) are connected by directed edges if they share one or more target-selective scaffolds. In this case, all scaffolds correspond to selective compounds, and the directionality of the edges indicates target (A over B) selectivity. In addition, the width of the edges is scaled according to the number of target-selective scaffolds. In Figure 3a, different selectivity patterns are observed. Figure S7 of the Supporting Information shows the corresponding networks for the 10- and 100-fold selectivity levels where similar observations can be made. As shown in Figure 3a, in addition to binary selectivity relationships, there are inverse relationships (where some scaffolds are selective for target A over B and others for B over A) and

also complex selectivity patterns. In addition, "selectivity hubs" become apparent, that is, individual targets with scaffold selectivity over several others. For example, the cluster formed by blue nodes at the upper left in Figure 3a represents selectivity relationships among the closely related serine proteases factor Xa (target ID 351), thrombin (352), and factor IXa (358) where multiple scaffolds generate compounds that are at least 50-fold selective for factor Xa over the other two proteases. Moreover, the cluster of pink nodes in the center corresponds to closely related dipeptidyl peptidases (DPPs) where single or multiple scaffolds are selective for DPP4 over related DPPs or pairs of DPPs. Figure 3b shows seven representative scaffolds that produce compounds selective for DDP4 over other DDPs and the selectivity relationships that they form. These scaffolds and the corresponding compounds are provided in Table S3 of the Supporting Information. Such scaffolds can be collected as starting points for generating compounds that are highly selective for a particular target over other closely related ones.

In summary, systematic mining of a publicly available compound data has revealed that small sets of target-selective scaffolds represented by multiple compounds exist, although selectivity data are sparsely distributed. These target-selective scaffolds are represented by up to 21 compounds that are highly selective for an individual target over one or two others. However, the majority of currently available target-selective scaffolds (at different selectivity levels) are only represented by individual compounds. Thus, many scaffolds are available for further experimental evaluation that might yield target-selective compounds. Importantly, however, selectivity patterns can be observed around specific targets that are formed by multiple target-selective scaffolds and establish different target relationships, which can also be exploited in the design of target-selective compounds.

EXPERIMENTAL PROCEDURES From BindingDB,[5] compounds with reported activity against human targets were extracted. If multiple potency measurements were reported in a BindingDB entry, their geometric mean was calculated as the final single potency value. Hierarchical scaffolds[4] were extracted from active compounds that represent ring systems and rings connected by linkers after removal of substituents. Compounds and scaffolds were represented in SMILES format[8] for processing. Network representations were generated with Cytoscape.[9] The method to determine target-selective scaffolds and the selectivity level assignments of scaffolds are detailed in the Methods of the Supporting Information. Scaffold and target selectivity analysis were carried out using in-house Pipeline Pilot[10] and Perl programs. These programs are described in the Methods of the Supporting Information and are available via the following URL: http://www.lifescienceinformatics.uni-bonn.de ("Downloads").

SUPPORTING INFORMATION AVAILABLE Details of scaffold selectivity analysis (Methods); tables reporting all targets investigated in this study (Table S1) and the scaffolds and corresponding compounds that are discussed (Tables S2 and S3); and figures presenting the annotated scaffold-based target network (Figure S1), the distribution of community-selective scaffolds (Figure S2), the selectivity-based scaffold classification scheme (Figure S3), the distribution of purely selective scaffolds (Figure S4), the methodology applied to identify scaffolds with exclusive target selectivity (Figure S5), the structures of target-selective scaffolds (Figure S6), and the network representations of selectivity patterns at different selectivity levels (Figure S7). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author: *To whom correspondence should be addressed. Tel: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

REFERENCES

(1) Evans, B. E.; Rittle, K. E.; Bock, M. G.; Dipardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S. Methods for Drug Discovery: Development of Potent, Selective, Orally Effective Cholecystokinin Antagonists. *J. Med. Chem.* **1988,** *31,* 2235–2246.

(2) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are Target-Family-Privileged Substructures Truly Privileged?. *J. Med. Chem.* **2006,** *39,* 2000–2009.

(3) Hu, Y.; Wassermann, A. M.; Lounkine, E.; Bajorath, J. Systematic Analysis of Public Domain Compound Potency Data Identifies Selective Molecular Scaffolds across Druggable Target Families. *J. Med. Chem.* **2010,** *53,* 752–758.

(4) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996,** *39,* 2887–2893.

(5) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* **2007,** *35,* D198–D201.

(6) PubChem. http://pubchem.ncbi.nlm.nih.gov/ (accessed September 1, 2009).

(7) Mestres, J.; Gregori-Puigjané, E.; Valverde, S.; Solé, R. Data completeness—The Achilles heel of drug-target networks. *Nat. Biotechnol.* **2008,** *26,* 983–984.

(8) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988,** *28,* 31–36.

(9) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003,** *13,* 2498–2504.

(10) *Scitegic Pipeline Pilot*, Student Edition, Version 6.1; Accelrys, Inc.: San Diego, CA, 2007.