Developing a Drug-like Natural Product Library[⊥]

Ronald J. Quinn,*[,†] Anthony R. Carroll,[†] Ngoc B. Pham,[†] Paul Baron,[†] Meredith E. Palframan,[†] Lekha Suraweera,[†] Gregory K. Pierens,[†] and Sorel Muresan[‡]

Eskitis Institute, Griffith University, Brisbane, Queensland, 4111, Australia, and AstraZeneca R&D Mölndal, Pepparedsleden 1, 431 83 Mölndal, Sweden

Received September 26, 2007

Addressing drug-like/lead-like properties of biologically active small molecules early in a lead generation program is the current paradigm within the drug discovery community. Lipinski's "rule of five" has become the most commonly used tool to assess the relationship between structures and drug-like properties. Sixty percent of the 126 140 unique compounds in *The Dictionary of Natural Products* had no violations of Lipinski's "rule of five". We have isolated 814 natural products based on their expected drug-like/lead-like properties to generate a natural product library (NPL) in which 85% of the isolated compounds had no Lipinski violations. The library demonstrates the feasibility of obtaining natural products known for rich chemical diversity with the required physicochemical properties for drug discovery. The knowledge generated in creation of the library of structurally characterized pure natural products may provide opportunities to front-load lead-like property space in natural product drug discovery programs.

While natural products represent a rich source of therapeutically useful compounds,[1–7] interest in the development of natural products by the pharmaceutical industry has declined. Natural products, however, remain biologically validated and, as such, should provide a good lead generation option. If natural products are to once again be of interest to the wider drug discovery world, they need to be able to be utilized within the lead generation paradigm.

We have investigated a biosynthetic enzyme/target correlation and established that an imprint of recognition of protein surfaces during biosynthesis is transferred to recognition of therapeutically useful enzyme targets. Comparison of X-ray crystallographic structures of the biosynthetic enzymes chalcone isomerase, chalcone synthase, and anthocyanidin synthase with three X-ray structure complexes of protein kinases with flavonoids identified a shared recognition. This recognition occurs at a level beyond the SCOP fold classification. Classification of proteins at the fold level according to the Structural Classification of Proteins (SCOP) database[8] is based upon the arrangement of major secondary structures and topological connections. Protein fold topology (PFT),[9] as exemplified by the flavonoid/kinase PFT, is defined as a similar arrangement of different secondary structures around the active site. A similar PFT at the target level has been demonstrated in a study of different fold targets of the same natural product.[10]

This biosynthetic enzyme/target correlation provides the underlying reason why natural products are validated starting points for drug design and explains the success of compound libraries based on natural product starting points.[11–13] It has been proposed that the large compound libraries used in HTS may not reflect the rich diversity of a smaller, purified compound natural products library.[14] There have been several comprehensive reviews of the types of libraries that have been developed inspired by natural products.[15–19] The latest review of libraries from natural product-like scaffolds encompasses over 50 reported libraries.[16]Some of the most recent libraries have been based upon (±)-vasicine,[20] 3-chloro-4-hydroxyphenylacetamide,[21] flavonoids,[22] the Amaryllidaceae alkaloids,[23] fused bicyclic acetals,[24] and spiroketal structures.[25]

Interestingly, a comparison by Feher and Schmidt[26] showed that, overall, natural products are more similar to drugs than compounds obtained from combinatorial synthesis. A large proportion of natural products are biologically active and have favorable ADME/T properties (absorption, distribution, metabolism, excretion, and toxicology), despite the fact that they often do not satisfy proposed "drug-likeness" criteria. Current thinking in the generation of drug leads embodies the concept of achieving high molecular diversity within the boundaries of reasonable drug-like properties.[27] Natural products, which possess biochemical specificity and occupy a larger chemical space than synthetic compounds, will become favorable as lead structures for drug discovery if they comply with drug-like/lead-like criteria. Building a physiochemically "tuned" natural product library in line with the lead generation paradigm is a necessary step to promote natural products to their full potential.

The lead generation paradigm requires that compounds conform to current understanding of lead-like properties and demands that natural products offer the same pathway through lead optimization. In order to understand how physicochemical properties can be front-loaded in natural product drug discovery, we have undertaken to develop a drug-like set of natural products. This is a first step and may subsequently allow this knowledge to be applied to obtain drug-like or lead-like extracts for screening. In this paper, we evaluate natural products from *The Dictionary of Natural Products* (DNP) and from our physiochemically "tuned" natural product library (NPL) for "drug-likeness".

Lipinski[28] has proposed a simple set of easily calculated properties, the so-called "rule of five", which have been derived from the 90th percentile of drug candidates that reached phase II clinical trials. It is an algorithm consisting of four rules in which many of the cutoff numbers are five or multiples of five, thus originating the rule's name. To be drug-like, a candidate should have less than five hydrogen bond donors (HBD), less than 10 hydrogen bond acceptors (HBA), a molecular weight of less than 500 Da, and a partition coefficient log *P* of less than 5. The aim of the "rule of five" is to highlight possible bioavailability problems if two or more properties are violated.

Other parameters also have been used to predict favorable DMPK (drug metabolism and pharmacokinetics) outcomes, such as rotatable bonds, polar surface area (PSA), log *D*, and counts of nitrogen and oxygen atoms.[29] It should be noted that natural products exhibit a wide range of flexibility, from rigid conformationally constrained molecules to very flexible compounds. Rigid molecules are normally missing from combinatorial libraries since synthesis of highly constrained molecules is generally more difficult.

[⊥] Dedicated to Dr. G. Robert Pettit of Arizona State University for his pioneering work on bioactive natural products.

* To whom correspondence should be addressed. Tel: +61 7 3735 6006. Fax: +61 7 3735 6001. E-mail: r.quinn@griffith.edu.au.

[†] Eskitis Institute, Griffith University.

[‡] AstraZeneca R&D Mölndal.

10.1021/np070526y CCC: $40.75
Published on Web 02/08/2008

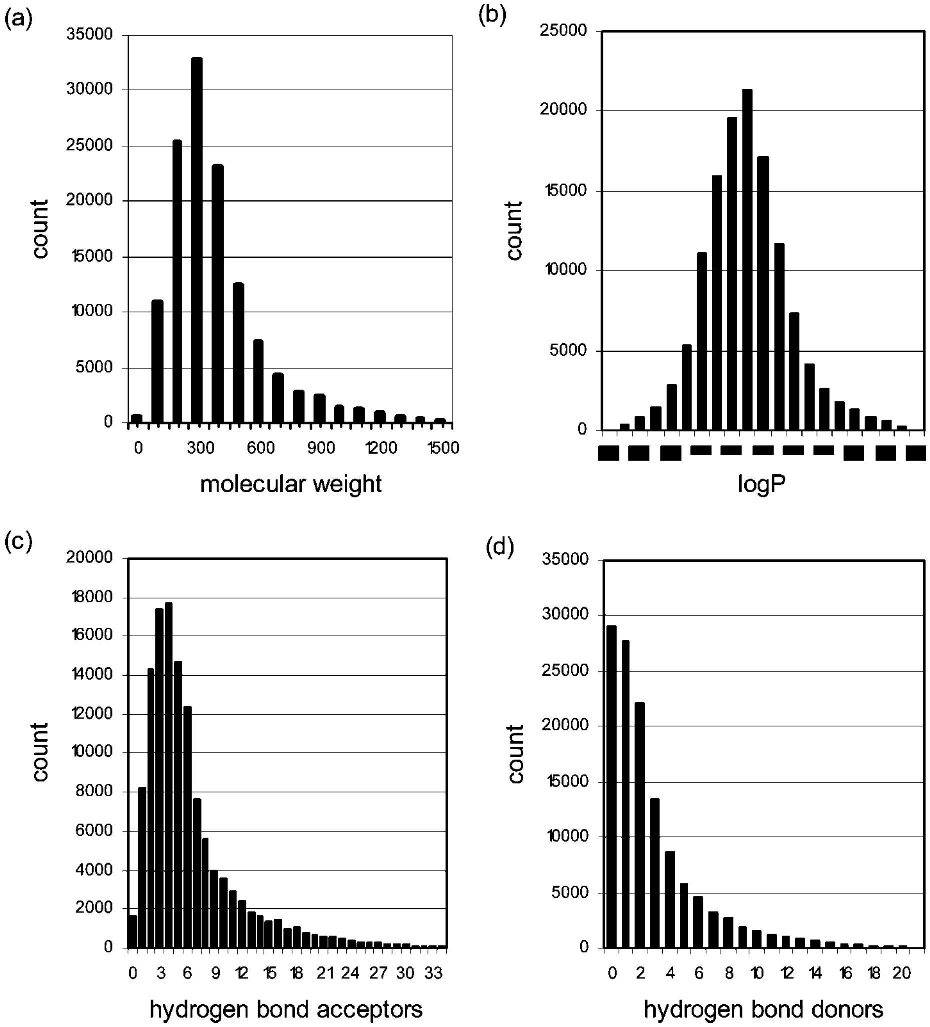

Figure 1. Histogram for DNP database (126 140 unique compounds) showing molecular weight, calculated log *P*, and hydrogen bond acceptors and donors.

Drug molecules are generally developed from less complex lead compounds. These lead compounds usually have a smaller number of rings, less rotatable bonds, and smaller molecular weight and are more hydrophilic.[30] Therefore rules for "lead-likeness"[31] have also been proposed, which have slightly more stringent criteria than the Lipinski's "rule of five" such as molecular weight < 450 and log $P < 4$. An even more restrictive set of rules have been defined for fragment screening.[32] Both these rules have been designed so that the compounds found can progress through traditional medicinal chemistry optimization.

The *Dictionary of Natural Products* (DNP)[33] is a comprehensive database of natural products containing ca. 171 000 entries derived from literature reports on the isolation and identification of compounds from diverse biota. The DNP compounds (ca. 171 000) were processed to neutralize salts and to remove duplicate compounds. The unique compounds (126 140) were saved in SMILES format. The resulting compounds were examined for Lipinski properties (molecular weight, calculated log *P*, hydrogen bond acceptors and donors).[34,35]

The four individual Lipinski properties were analyzed, and the histograms for molecular weight, calculated log *P*, and hydrogen bond acceptors and donors are shown in Figure 1. The histograms were expressed as counts to highlight the actual numbers of compounds in each binned plot.

The histogram of molecular weight (Figure 1a) showed very similar distribution to that reported by Feher and Schmidt.[26] The molecular weight distribution was a maximum at 300–400 Da, and about 26% of the compounds analyzed had molecular weights over 500 Da.

The histogram of calculated log *P* (Figure 1b) showed a Gaussian distribution with a maximum at 2–2.5 log *P* units. There were some compounds with very large calculated log *P*'s. This was probably because the training database/algorithm used to calculate log *P* may not suit the types and combinations of functional groups found in natural products. It is interesting to note that we have experimentally measured the log *P*'s for many of our natural product compounds by a HPLC method[36] and have found that the calculated and experimental values can vary by up to 12 units (unpublished result). These large variations were most noticeable in compounds containing bromine. The predicted log *P* values for these compounds were generally very large. Compounds that can hide functional groups from solvents, such as cyclic peptides, had very negative calculated log *P* data, although the experimental log *P* values were in the range 1–3. Even with the shortcomings of these calculations, 81% of the DNP data set had a calculated log *P* of less than 5.

The distribution of the hydrogen bond acceptors (HBA) (Figure 1c) peaked at about 3–5 and fell off very quickly to a maximum of 34. About 18% of the database had over 10 acceptors. The histogram of hydrogen bond donors (HBD) (Figure 1d) showed a steeply decreasing function from 0 (peak of the distribution) to a maximum of 31. About 20% of the compounds had donors greater than 5.

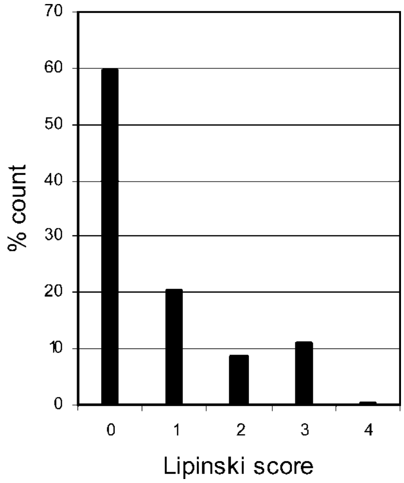

Figure 2. Histogram of Lipinski violations as a percentage of the DNP data set.

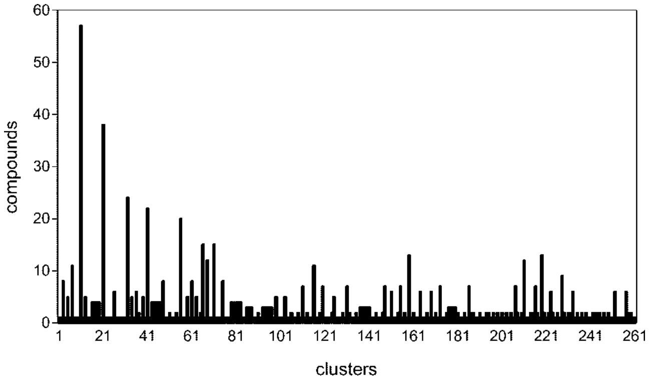

Figure 3. Clusters of NPL using a Tanimoto distance of 0.3.

Our data also agreed with the work reported by Feher and Schmidt[26] with a smaller natural product database (3287 compounds). About 80% of the compounds had less than two violations of Lipinski's "rule of five" (Figure 2).

We have established a natural product library (NPL) consisting of pure structurally elucidated natural products. The NPL was designed to obtain compounds with improved drug-like/lead-like properties. Natural products with no Lipinski violation were chosen from the DNP. Plants from Queensland, mainland China, and Papua New Guinea in our biota collection were analyzed by electrospray ionization mass spectrometry coupled with a nitrogen detector to determine if they contained the targeted natural products. This analysis indicated that 1000 plants in the collection might contain known natural products with no Lipinski violations. Compounds were isolated by mass-directed purification and structures determined by NMR spectroscopy. The data for the NMR analysis were acquired using a cold probe attached to a Varian Unity Inova 600 MHz spectrometer (COSY, HSQC, and HMBC). While biota were selected based on mass spectrometric evidence of containing known compounds, 60% of the isolated compounds were published natural products with a further 39% of unpublished analogues and 1% of novel compounds. Altogether 814 unique compounds were isolated from the 1000 biota. The structures of the isolated compounds were clustered using a Tanimoto distance of 0.3 and displayed a high level of structural diversity. The cluster set showed 261 clusters with 129 singletons, 56 doublets, and 17 triplets. The maximum number of compounds in a cluster was 57 (Figure 3).

The majority of compounds were isolated in sufficient quantities for thorough biological evaluation to be undertaken in the future (10–50 mg). Most of these compounds were isolated in yields ranging from 0.01 to 0.5% dry weight, and about 1% of the purified compounds were isolated in a yield of 0.001% dry weight.

The structures of the compounds in NPL (814 compounds) were converted into SMILES format and submitted for analysis for their Lipinski properties.[34,35] It was found that 95% of the compounds had less than two violations, with 85% having no violations. These values were improved over the DNP data set, which had 80% and 60% in these categories, respectively. A comparison of the distributions for the individual parameters for the DNP and NPL is shown in Figure 4. The histograms show only the data within the Lipinski interested regions (molecular weight < 500, $-2 < \log P < 5$, HBA < 10, and HBD < 5) and are expressed as a percentage count of their respective databases. In all cases the distributions of the NPL were enhanced for the Lipinski properties (peaks of the distributions moved to more lead-like properties) when compared to the DNP.

The molecular weight distribution (Figure 4a) peaks at 300–400 Da for both the DNP and the NPL. Above this the percentages were reduced for the NPL with respect to the DNP, and below this the percentages values were enhanced for the NPL with respect to the DNP. This improved profile for molecular weight is exactly what is desirable for a more lead-like library. The proportions of the two databases that satisfy Lipinski's molecular weight property (<500 Da) were 73% for DNP and 91% for NPL.

The distribution maximum of calculated log *P* (Figure 4b) for the NPL is in the same region as the DNP (log *P* of 2–3). There was a trend similar to that for the molecular weight distribution. The percentage of compounds, with calculated log *P* greater than 3 for the NPL, was reduced as compared to DNP, while the percentage of compounds with a calculated log *P* less than 2 was enhanced. The NPL relative enhancement of this Lipinski property was 9% compared to the DNP. Notably in the NPL, 87% of compounds satisfied the rule for log *P*.

The distribution of the histogram of hydrogen bond acceptors (HBA) (Figure 4c) for the NPL showed a maximum at 4–5 acceptors. The number of acceptors below 3 was reduced; those between 3 and 7 increased and above 8 reduced for the NPL as compared to the DNP values. Therefore the compounds in the NPL had enhanced the acceptable number of acceptors to 93% compared to 82% in the DNP. This is probably due to the NPL having an enriched percentage of N-containing compounds. The peak of the distribution for the hydrogen bond donors (HBD) for the NPL is at 1 with a significant increase in 1 or 2 donors as compared to the DNP (Figure 4d). The overall enhancement over the DNP values for this HBD property was 13%.

The overall summary of the four Lipinski parameters for the two databases is shown in Figure 5. The targeted natural product library has an average enhancement on drug-like properties of 13%, with the maximum enhancement for the molecular weight (18%) and the minimum enhancement for calculated log *P* (9%).

While 60% of the 126 140 unique compounds in *The Dictionary of Natural Products* had no violations of Lipinski's "rule of five", 85% of the isolated compounds had no violations. The library of isolated natural products could be classed as lead-like. This indicates that there should be an increase in success in finding a lead-like molecule with improved DMPK properties within a library such as NPL. We are annotating the isolated compounds with respect to their cellular phenotypes in a forward chemical genetics approach. In a wider context, the knowledge generated in creation of the library of structurally characterized pure natural products may provide opportunities to front-load lead-like property space in natural product drug discovery programs.

Experimental Section

Calculation of Physicochemical Properties. The natural product compounds were taken from Chapman and Hall's *Dictionary of Natural Products* (DNP, April 2005).[33] The DNP compounds (ca. 171 000) were processed to neutralize salts and to remove duplicate compounds. The

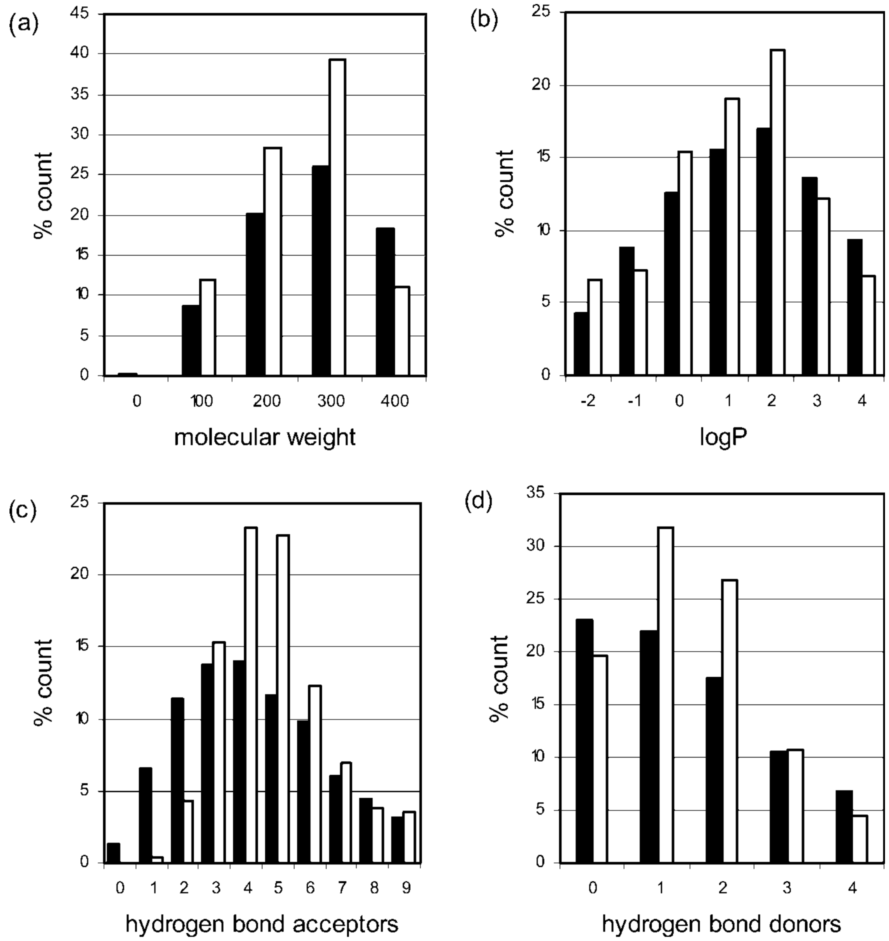

Figure 4. Comparison of property distribution for the two data sets (DNP (black), NPL (white)). *x*-Axis label is the lower limit of binned data, e.g., 100 is equivalent to 100–200.

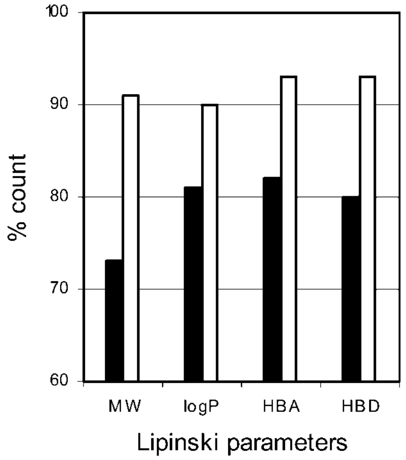

Figure 5. Comparison of content of the two databases that obey Lipinski's "rule of five" [DNP (black), NPL (white)].

unique compounds (126 140) were saved in SMILES format. The purified lead-like/drug-like natural product library (NPL, 814 compounds) was converted into SMILES format. Both libraries (DNP, NPL) were examined for Lipinski properties (molecular weight, calculated log *P*, hydrogen bond acceptors and donors).[34,35]

Natural Product Isolation and Structure Elucidation. A subset of 1000 plant biota from NPD biota library was selected for this study. The biota were collected from Queensland, the People's Republic of China, and Papua New Guinea and were analyzed in a small scale using an electrospray time-of-flight mass spectrometer coupled with a nitrogen detector. The compounds were isolated by mass directed purification using an Agilent 1100 Series LC/MSD LC-MS. NMR data were acquired using a cold probe attached to a Varian INOVA 600 MHz NMR spectrometer.

Acknowledgment. We thank AstraZeneca for financial support.

References and Notes

(1) Cragg, G. M.; Newman, D. J.; Snader, K. M. *J. Nat. Prod.* **1997**, *60*, 52–60.
(2) Tan, D. S.; Foley, M. A.; Shair, M. D.; Schreiber, S. L. *J. Am. Chem. Soc.* **1998**, *120*, 8565–8566.
(3) Rouhi, A. M. *Chem. Eng. News* **2003**, *81*, 104–107.
(4) Newman, D. J.; Cragg, G. M.; Snader, K. M. *J. Nat. Prod.* **2003**, *66*, 1022–1037.
(5) Newman, D. J.; Cragg, G. M. *J. Nat. Prod.* **2004**, *67*, 1216–1238.
(6) Lam, K. S. *Trends Microbiol.* **2007**, *15*, 279–289.
(7) Newman, D. J.; Cragg, G. M. *J. Nat. Prod.* **2007**, *70*, 461–477.
(8) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. *J. Mol. Biol.* **1995**, *247*, 536–540.
(9) McArdle, B. M.; Campitelli, M. R.; Quinn, R. J. *J. Nat. Prod.* **2006**, *69*, 14–17.
(10) McArdle, B. M.; Quinn, R. J. *ChemBioChem* **2007**, *8*, 788–798.
(11) Koch, M. A.; Waldmann, H. *Drug Discovery Today* **2005**, *10*, 471–483.
(12) Koch, M. A.; Wittenberg, L.-O.; Basu, S.; Jeyaraj, D. A.; Gourzoulidou, E.; Reinecke, K.; Odermatt, A.; Waldmann, H. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 16721–16726.
(13) Breinbauer, R.; Vetter, I. R.; Waldmann, H. *Angew. Chem., Int. Ed. Engl.* **2002**, *41*, 2879–2890.
(14) Dobson, C. M. *Nature (London)* **2004**, *432*, 824–828.
(15) Abreu, P. M.; Branco, P. S. *J. Braz. Chem. Soc.* **2003**, *14*, 675–712.
(16) Boldi, A. M. *Curr. Opin. Chem. Biol.* **2004**, *8*, 281–286.
(17) Lee, M. L.; Schneider, G. *J. Comb. Chem.* **2001**, *3*, 284–289.
(18) Abel, U.; Koch, C.; Speitling, M.; Hansske, F. G. *Curr. Opin. Chem. Biol.* **2002**, *6*, 453–458.
(19) Messer, R.; Fuhrer, C. A.; Haener, R. *Curr. Opin. Chem. Biol.* **2005**, *9*, 259–265.

(20) Shevyakov, S. V.; Davydova, O. I.; Pershin, D. G.; Krasavin, M.; Kravchenko, D. V.; Kiselyov, A.; Tkachenko, S. E.; Ivachtchenko, A. V. *Nat. Prod. Res., Part A* **2006**, *20*, 735–741.

(21) Davis, R. A.; Pierens, G. K.; Parsons, P. G. *Magn. Reson. Chem.* **2007**, *45*, 442–445.

(22) Yao, N.; Song, A.; Wang, X.; Dixon, S.; Lam, K. S. *J. Comb. Chem.* **2007**, *9*, 668–676.

(23) Keaney, G. F.; Johannes, C. W. *Tetrahedron Lett.* **2007**, *48*, 5411–5413.

(24) Milroy, L.-G.; Zinzalla, G.; Prencipe, G.; Michel, P.; Ley, S. V.; Gunaratnam, M.; Beltran, M.; Neidle, S. *Angew. Chem., Int. Ed.* **2007**, *46*, 2493–2496.

(25) Zinzalla, G.; Milroy, L.-G.; Ley, S. V. *Org. Biomol. Chem.* **2006**, *4*, 1977–2002.

(26) Feher, M.; Schmidt, J. M. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218–227.

(27) Koehn, F. E.; Carter, G. T. *Nat. Rev. Drug Discovery* **2005**, *4*, 206–220.

(28) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

(29) Proudfoot, J. R. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1647–1650.

(30) Hann, M. M.; Leach, A. R.; Harper, G. *J. Chem. Inf.Comput. Sci.* **2001**, *41*, 856–864.

(31) Hann, M. M.; Oprea, T. I. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255–263.

(32) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. *Drug Discovery Today* **2003**, *8*, 876–877.

(33) *Dictionary of Natural Products on CD-Rom*; Chapman and Hall/CRC Press: London, 2005.

(34) *DaylightToolkit, version 4.81*; Daylight Chemical Information Systems Inc.: Aliso Viejo, CA.

(35) *ACD/LogP, version 9*; Advanced Chemistry Development Inc.: Toronto, Ontario, Canada, 2005.

(36) Lombardo, F.; Shalaeva, M. Y.; Tupper, K. A.; Gao, F.; Abraham, M. H. *J. Med. Chem.* **2000**, *43*, 2922–2928.

NP070526Y