Chemical substructures that enrich for biological activity

Justin Klekota\textsuperscript{1,2} and Frederick P. Roth\textsuperscript{3,4,*}

\textsuperscript{1}Harvard University Graduate Biophysics Program, \textsuperscript{2}Harvard Medical School, 250 Longwood Avenue, \textsuperscript{3}Harvard Institute of Chemistry and Cell Biology, \textsuperscript{4}Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue and \textsuperscript{4}Center for Cancer Systems Biology, Dana-Farber Cancer Institute, One Jimmy Fund Way, Boston, MA 02115, USA

Received on May 6, 2008; revised on August 13, 2008; accepted on September 7, 2008
Advance Access publication September 10, 2008
Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Certain chemical substructures are present in many drugs. This has led to the claim of ‘privileged’ substructures which are predisposed to bioactivity. Because bias in screening library construction could explain this phenomenon, the existence of privilege has been controversial.

Results: Using diverse phenotypic assays, we defined bioactivity for multiple compound libraries. Many substructures were associated with bioactivity even after accounting for substructure prevalence in the library, thus validating the privileged substructure concept. Determinations of privilege were confirmed in independent assays and libraries. Our analysis also revealed ‘underprivileged’ substructures and ‘conditional privilege’—rules relating combinations of substructure to bioactivity. Most previously reported substructures have been flat aromatic ring systems. Although we validated such substructures, we also identified three-dimensional privileged substructures. Most privileged substructures display a wide variety of substituents suggesting an entropic mechanism of privilege. Compounds containing privileged substructures had a doubled rate of bioactivity, suggesting practical consequences for pharmaceutical discovery.

Contact: fritz_roth@hms.harvard.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The observation that commercially available drugs have physical properties that distinguish them from other compounds led to the establishment of Lipinski’s ‘Rule of 5’ to predict drug absorption and permeation (Lipinski et al., 2001). While the likelihood that any given compound which satisfies this rule will become a drug remains small, the Rule of 5 has been a valuable guide for the design of chemical libraries. The need exists to further enrich chemical libraries with potential drug molecules.

The abundance of certain substructures in commercially available drugs has motivated the search for privileged substructures, i.e. substructures associated with biological activity (DeSimone et al., 2004; Horton et al., 2003). Among the reported privileged substructures are benzodiazepines (Evans et al., 1988) and hydrophobic ring systems (Ariens et al., 1979, Fig. 1A) which are present in drugs active against various protein targets (Andrews and Lloyd, 1982).

The shape of privileged substructures may be preferred by hydrophobic pockets on protein surfaces (Bondensgaard et al., 2004; Hajduk et al., 2000; McGaughey et al., 1998), or have structural homology to biological substrates (Fig. 1B; Jacobson, 2001; McGaughey et al., 1998). For example, the benzodiazepine scaffold (Evans et al., 1988) (1) in drug compounds (9) (Fig. 1C) is active against opioid receptors and other protein targets (Marsters, 1994; Patchett and Nagrund, 2000). This may be explained by structural homology to endogenous biomolecules (8) (Sangameswaran et al., 1986) or to peptide β-turns (Ripka et al., 1993). Similarly, Indole (2) and Purine (3), present in many drug compounds (9, 10) (DeSimone et al., 2004; Dinnell et al., 2001; Heinelt et al., 2001; Jacobson, 2001; Willoughby et al., 2002), are also present in endogenous biomolecules such as tryptophan (7) and ATP (6).

While many of the reported privileged substructures are flat, aromatic ring systems, there are exceptions. For example, spiroperiderines (Klabunde et al., 2002; Patchett and Nagrund, 2000; Patchett et al., 1995) (4, 11) and cyclic peptides (seen in cyclosporin A, for example) (Horton et al., 2002) have three-dimensional geometries.

The activity of privileged substructures is distinct from that of promiscuous inhibitors which act by molecular aggregation and inhibit proteins non-selectively (McGovern et al., 2002). Many compounds containing privileged substructures bind proteins selectively, as revealed by NMR studies of biphenyl (5), for example (Fig. 1A; Hajduk et al., 2000), also present in the drug diflunisal (12) (Fig. 1C).

Privileged substructures remain controversial because their abundance in drug compounds may be a trivial consequence of their abundance in chemical libraries (DeSimone et al., 2004). Numerous computational analyses have identified privileged substructures abundant in bioactive compounds (Bemis and Murcko, 1996; Lewell et al., 1998; Nilsson et al., 2001; Sheridan, 2003; Wagener and van Geresteen, 2000), without considering bias towards some substructures in compound library construction. In contrast, others have employed decision trees (Rusinko et al., 1999, 2002; van Rhee, 2003; Young and Hawkins, 1995) to identify substructures that discriminate activity from inactivity within a given collection of compounds. The decision tree estimates the conditional probability of activity given the combination of substructures present (or absent) in a compound, while accounting for the abundance of substructures
growth inhibition in 70 cancer cell lines. Compounds in the NCI library scoring in the top 3% of at least one assay were identified as active (11.4% of all compounds); the more liberal assay threshold of 3% was used because compounds meeting this threshold were active below the reported Chembridge screening concentration of 10 µM. Because most of the compounds in the Chembridge library had molecular weights below 500, only compounds with a molecular weight below 500 were examined in the NCI library. In addition, we studied an assay of the Chembridge Microformat library measuring inhibition of histone acetylation and two assays of the Chembridge Diverse Set E library (not included above) measuring arginine methyltransferase inhibition (Cheng et al., 2004) and selective killing of Neu-overexpressing cells (Fantin et al., 2002).

### 2.2 Fragmentation of compounds

The structure of each compound was converted to a SMILES string using Daylight’s mol2smi algorithm. A collection of 4860 unique substructures were generated by fragmenting each compound SMILES in the Chembridge Diverse Set E library using the Daylight SMARTS and SMIRKS toolkits and one of six fragmentation strategies, including RECAP (Csizmadia, 2000; Lowell et al., 1998, see Supplementary Material) which employs retrosynthetic analysis and tends to produce substructures that would be useful in guiding medicinal chemistry optimization and combinatorial library design. Each fragmentation strategy was applied exhaustively using a series of virtual chemical reactions (represented by SMIRKS strings). The resulting substructures were represented as SMARTS strings (symbols representing the non-hydrogen wildcard ‘[*]’ were used where appropriate). The generated substructures were then pooled and those appearing at least five times in the library were retained, yielding a non-redundant set of 4860 chemical substructures. We note the analogy of this strategy to previous graph mining methods (Cook and Holder, 2006; Nijssen and Kok, 2004; Rosenkranz and Klopman, 1990). The Daylight SMARTS toolkit was applied to generate an array of 1’s and 0’s indicating the presence or absence of each substructure in each molecule in each library.

### 2.3 Decision trees

In order to find the substructures most associated with biological activity, decision trees (Rusinko et al., 1999, 2002; van Rhee, 2003; Young and Hawkins, 1995) were used to partition compounds in the Chembridge Diverse Set E library based on the presence or absence of highly discriminating substructures chosen from the set of 4860 substructures. The most discriminating substructures were identified based on mutual information between substructure presence and compound activity. Compounds were then partitioned into subgroups depending on the presence or absence of a given discriminating substructure, and those subgroups were further partitioned (recursively) based on additional discriminating substructures. To avoid overfitting, discriminating substructures were chosen using the Bayesian information criterion (BIC) (Friedman and Goldszmidt, 1996; King et al., 2003, see Supplementary Material). The final partitions (“leaf nodes” in the decision tree) represent sets of compounds with (or without) specific substructures that are enriched or depleted in activity relative to other compounds in the library.

### 2.4 Significance of substructure activity

Each discriminating substructure selected by the decision tree was tested for association with activity within each individual assay and using the definition of activity based on multiple assays. The cumulative hypergeometric test of association was used (Klekota et al., 2005, see Supplementary Material). For each substructure, the number of individual assays yielding significant ($P < 0.01$) associations (either positive or negative) was counted.

### 2.5 Ability of the decision tree to enrich for bioactivity

The ability of the decision tree to enrich for compound activity in various assays and compound libraries was determined. Compounds assigned to
We examined 24 cell-based phenotypic assays applied to the 16 320-compound Chembridge Diverse Set E library. These assays encompass a variety of chemical-induced phenotypes including mitotic arrest, endocytosis inhibition and histone acetylation (Boyce et al., 2005; Feng et al., 2003; Haggarty et al., 2000, 2003; Mayer et al., 1999; Nieland et al., 2002; Yarrow et al., 2003, 2005). This set of assays was selected from an original set of 85 assays, excluding assays with phenotypes attributable to compound fluorescence, toxicity (cell-death), non-specific transcriptional upregulation or inhibition of luciferase (a commonly used assay reporter). Because false positive ‘promiscuous inhibitors’ form molecular aggregates that are less membrane permeable (McGovern et al., 2002), we used only cell-based assays.

We focused primarily on the Chembridge Diverse Set E library, since it is not highly biased towards particular protein target classes. We confirmed its diversity, showing that the average pairwise Tanimoto coefficient—a measure of chemical substructure similarity—is 0.2 within this library. This is well below the threshold of 0.85 that is widely used to classify compounds as similar.

The set of 16 320 Chembridge Diverse Set E compounds was partitioned using a decision tree (Fig. 2). In this tree, the ‘root’ node corresponds to the set of all compounds. Compounds were successively divided into ever smaller subsets according to the presence or absence of ‘discriminating’ substructures. The ‘X’ symbol indicates a non-hydrogen atom and all hydrogen atoms in ‘X’-containing substructures (whether shown or implied) must be exactly matched. All other substructures have unspecified patterns of hydrogen and non-hydrogen atom substitution. At each node, the discriminating substructure that was used to divide the corresponding set of compounds was chosen using an unbiased information-theoretic criterion (see Section 2). Each compound was ultimately classified into one of 44 ‘leaf nodes’—compound subsets that are enriched or depleted in biological activity relative to the rest of the library—based on the presence or absence of 43 ‘discriminating’ substructures (Fig. 2).

Discriminating substructures selected by the decision tree (Fig. 2 and Supplementary Fig. S1) include many which were reported as privileged (DeSimone et al., 2004; Horton et al., 2003). For example, indole is associated with an increase in biological activity in the Chembridge assay among compounds lacking the substructures shown at nodes 1, 3, 5, 8 and 11. Interestingly, the selected indole substructure (node 16) had multiple non-hydrogen atoms (‘X’) attached to it, supporting previous intuition that privileged substructures may represent molecular scaffolds enriched for favorable binding entropy rather than enthalpy or complementary charge (Bondensgaard et al., 2004; Hajduk et al., 2000; Jacobson, 2001; McGaughey et al., 1998). Other potential scaffolds with multiple non-hydrogen substituents were also associated with activity: these include pyrrole (substructure at node 23) and benzene (substructures at nodes 32 and 39), which are components of indole and certain amino acids. Quinoline (13) (Fig. 3A) with an attached hydroxyl group (substructure at node 43) was also associated with increased activity. This substructure resembles the reportedly privileged substructures quinoxaline (14) (Fig. 3A) and quinazoline (15) (Fig. 3A) (Horton et al., 2003).

(20) is structurally similar to naphthoquinone (18) and is present in a variety of biologically active compounds including the FDA-approved anti-coagulant phenindione (20) and pesticides (Braselton et al., 1992) (Fig. 3B). 1,3-indandione has homology to a structural component of vitamins K1 and K2 and competes with vitamin K binding (Mount and Feldman, 1983).

In contrast to reportedly privileged substructures, some of the discriminating substructures associated with bioactivity are neither flat nor aromatic (Fig. 2 and Fig. S1; substructures at nodes 7 and 35). These substructures contain rings and double bonds that contribute to rigidity (a feature of many reported privileged substructures), but they also have sp³-hydridized atoms that make them richer in three-dimensional geometry. Notably, compounds in the Chembridge Diverse Set E library containing the substructure at node 7 are structurally homologous to three compounds present in the NCI library—NSC636679 (21), NSC634791 (22) and NSC618757 (23) (Fig. 3C). These compounds have been reported to inhibit cancer cell growth through inhibition of ABCB1 (MDR1), a membrane transport protein implicated in multi-drug resistance of cancers (Szakács et al., 2004). Each of the compounds contain the substructure at node 7 flanked by aromatic rings, forming a symmetric molecule. Examination of the ABCB1 structure (Seigneuret and Garnier-Suillerot, 2003) suggests that molecules may interact with ABCB1’s two ATP-binding sites. This observation motivates the exploration of other molecular scaffolds with rich three-dimensional geometries and symmetries.
Some substructures (notably the substructures at nodes 10, 24 and 27) are ‘underprivileged’, i.e. associated with decreased biological activity. Among these substructures are long carbon chains and chains of other sp³-hybridized atoms which are highly flexible and therefore likely increase the entropic cost of protein binding, in contrast to the privileged ring systems, which have less flexibility and predictably smaller entropic barriers to binding drug pockets: substructures only containing sp³-hybridized carbons were enriched in significantly fewer assays than substructures only containing aromatic carbons (P = 0.0006, see Supplementary Fig. S2) and substructures lacking rings were enriched in significantly fewer assays than substructures containing rings (P = 0.0036, see Supplementary Fig. S2). Underprivileged substructures should not necessarily be excluded from chemical libraries, as they may provide binding specificity, an important property of successful drugs; in fact, certain substructures (notably those at nodes 37 and...
and 40) become favorable in the presence of other substructures as depicted by the tree, but are not favored when considered individually (Table S1), demonstrating the ‘conditional’ privilege of certain substructures. Different aliphatic carbon chains distinguish the various K vitamins suggesting physiological significance of these substructures (Fig. 3B). It is also interesting that twice as many substructures were found to be enriched in activity than were found to be deficient.

The most discriminating substructure (node 1, the root node of the tree) represents the entire active portion (having sigma-orbital electron pairs) of the known metal chelator NIH (24) (Darnell and Richardson, 1999; Le and Richardson, 2004; Liang and Richardson, 2003) (Fig. 3D). This substructure is associated with a significant increase in biological activity. Metal chelators are reported to promote transcription non-specifically (Darnell and Richardson, 1999; Le and Richardson, 2004; Liang and Richardson, 2003) and inhibit other proteins affecting cell growth (Darnell and Richardson, 1999), so that this substructure is expected to correlate with bioactivity; many metal chelators were observed to confound our excluded gene reporter assays of the Chembridge Diverse Set E library (Randy King, HMS, personal communication).

To provide intuition about the compound sets corresponding to nodes in the tree, Supplementary Figure S3 shows representative structures for each leaf node of the tree shown in Figure 2 and Supplementary Fig. S1.

A separate tree generated using the same assay data with a fold-change threshold for activity yielded similar discriminating substructures (data not shown). Interestingly, the new tree included an additional three-dimensional substructure—a tricyclic ring system with a seven-membered ring resembling that of benzodiazepine (see Supplementary Fig. S4). The privileged status of this ring system was not evident in the original tree (perhaps due to the selection of two of its component rings in the original tree).

We explored an alternative definition of activity, requiring compounds to score in two or more assays. A tree trained using this definition contained many of the substructures present in the original tree and was similarly predictive of activity in independent assays despite the exclusion of 75% of compounds defined as active under the more permissive definition of activity (Fig. S9).

3.2 Substructures associated with general bioactivity

The significance of each substructure’s enrichment or depletion for activity in the decision tree was also confirmed by a statistical test of association ($\alpha=0.01$). Here, we tested the association of each substructure with activity in each individual assay. Because this test considered all compounds, significant associations indicate ‘unconditionally’ privileged (or underprivileged) substructures. An expanded set of 59 assays was examined (Table S1), including additional pure protein- or cell-extract-based assays and cellular toxicity assays, in addition to the 24 assays used to develop the decision tree. Substructures showing significant enrichment or depletion ($\alpha=0.05$) in three or more assays were considered broadly enriched. Nearly all discriminating substructures selected for use in the decision tree were corroborated by significant enrichment or depletion (the enrichment of many tree nodes in bioactivity exceeded the frequency of actives in our training set, Table S2). For example, the Indole scaffold (substructure at node 16) was enriched in 18 individual assays—consistent with its role as enriching for activity in the decision tree. There were however, a number of exceptions, e.g. substructures at nodes 9, 14, 37 and 40. These substructures, selected as discriminating in the presence of other substructures (marked ‘Combo’ in Table S1), did not show enhanced activity in individual assays or tended to be depleted in activity. These substructures are conditionally privileged, i.e. are associated with heightened activity only in the presence of other substructures. Thus, the decision tree reveals ‘rules of privilege’ that associate biological activity with specific combinations of substructures.

Given that the decision tree selected only 43 substructures out of 4860 available, many other privileged substructures are likely present in the Chembridge Diverse Set E library. Some of these unselected substructures may not have been selected due to their similarity to substructures already selected. Examination of recent chemical literature (DeSimone et al., 2004; Horton et al., 2003) identified additional substructures previously reported to be ‘privileged’ in the Chembridge Diverse Set E library. We found that many of these are indeed broadly enriched (as defined above) (Fig. S5), including biphenyl (25), 1,4-dihydropyridine (27), chromone (31), quinoxaline (33), indole (35) and benzimidazole (36).

3.3 The ability of discriminating substructures to enrich for bioactivity

We wondered whether these substructures in combination could enrich for activity to an extent that would be practically useful. To examine this question, we trained a decision tree using a randomly
selected 90% of Chembridge Diverse Set E library compounds, reserving the remaining 10% of compounds for testing. The resulting tree contained 39 discriminating substructures. Comparing these to the 43 discriminating substructures in the original tree, we found that 31 were identical, three differed only by a few explicit hydrogen atoms and the remaining five were structurally homologous. Each compound in the test set was mapped to a leaf node in the new tree based on substructure composition and assigned the corresponding activity score. Ranking compounds by activity score (Fig. 4A) revealed a substantial enrichment for active compounds; strikingly, there were 2–5 times more active compounds amongst the top 240-scoring compounds than among randomly chosen compounds. Similar results were obtained from a tree trained on only 50% of the data.

We wondered whether a decision tree trained on one set of assays to enrich for activity would be practically useful when applied to an independent set of assays. To this end, we labeled compounds in the Chembridge Diverse Set E library as active or inactive based on an assay measuring arginine methyltransferase inhibition (Cheng et al., 2004), which was not included in the original training set of 24 assays. Among the top-ranked 1200 compounds, the frequency of active compounds was 1.5–3 times greater than among randomly ranked compounds (Supplementary Fig. S6). Another assay not used in training, the extent of killing of Neu-overexpressing ‘oncogenic’ cells (Fantin et al., 2002) relative to wild-type cells, yielded a frequency of activity among top-ranked compounds that was 1.5–4 times higher than random compounds (Supplementary Fig. S7). Thus, the rules of privilege learned from one set of assays can be generally applied to substantially enrich for independent biological activities.

Because chemical libraries vary in their substructure composition, we wondered whether the rules of privilege learned from one chemical library would apply to independently constructed chemical libraries. To examine this question, we examined an assay of inhibition of histone acetylation in the Chembridge Microformat library. (This assay was among those applied to the Chembridge Diverse Set E library, which were used to train the original decision tree.) Each Microformat compound was mapped to a leaf node in the original decision tree (Fig. 2) and assigned the corresponding activity score. Amongst the top-ranked 450 compounds, the frequency of activity was 1.5–4 times higher than that of randomly chosen compounds (Supplementary Fig. S8). Repeating all of the above analyses after first removing compounds containing the substructure at node 1, the suspected metal chelators produced similar results. Thus, the rules of privilege determined from one chemical library allow substantial enrichment for activity within independent chemical libraries.

We wondered whether rules of privilege also had the power to enrich for bioactivity when both the chemical library and biological assays were independent of those used to train the decision tree (Fig. 2). Compounds of an NCI compound library were examined and identified as active if they scored in at least one assay measuring cancer cell line growth inhibition. Each NCI compound was mapped to a leaf node in the original decision tree (Fig. 2) and assigned the corresponding activity score. Amongst the top 1000 compounds, the frequency of activity was 1.5–3 times higher than that of randomly chosen compounds (Fig. 4B). This finding validates the concept of privileged substructure and shows that substructure properties learned from one dataset may be applied generally to multiple independent chemical libraries and bioactivities.

Activity of large NCI compounds (molecular weight >500) was poorly predicted based on substructures trained on the Chembridge library. The contributions of privileged substructures that we identified in a low-molecular weight library are likely to be diluted in larger molecules; furthermore, the mechanism of action of compounds with large molecular weights is likely to be qualitatively different from that of smaller compounds. Although privileged substructures may well exist among higher molecular weight compounds, these may need to be learned from a similar analysis applied to diverse biological assays of high-molecular weight compounds.

4 DISCUSSION

Our results validate the concept of privileged substructures, showing that many privileged substructures remain even after accounting for their overall abundance in the screened library. Moreover, privileged substructures identified as enriched for bioactivity in one library were also enriched for bioactivity within independent
chemical libraries and assays were not used to learn rules of privilege. We confirmed several previously reported privileged substructures. We also identified 'underprivileged' substructures depleted in biological activity, e.g. long chains of sp3-hybridized atoms. While previously reported privileged substructures have had flat aromatic ring systems, we identified privileged substructures with three-dimensional geometries and others may be found in the analysis of libraries containing more three-dimensional substructure. Furthermore, the observation that privileged scaffolds contain diverse substituents suggests the broad activity associated with privileged substructures is the result of favorable scaffold entropy, while activity against a given target is determined by entropic contributions in combination with complementarity of shape and charge resulting from enthalpic contributions of substituents. For many assays, prioritization of compounds based on substructure double the frequency of active compounds. Therefore, the use of 'rules of privilege' to design new chemical libraries with a preference for particular combinations of substructure could have important implications for pharmaceutical discovery.

ACKNOWLEDGEMENTS

We thank Stuart Schreiber of Harvard University for his support. We thank John Tallarico, Nathaniel Gray, Gabriel Berriz and Lan Zhang for their advice. We thank John Sullivan, Caroline Shamu and Su Chiang for their informatics assistance. We also thank the Harvard Biophysics Program for supporting this research.

Funding: Keck Foundation (to F.P.R.); National Institutes of Health (grants R01 HG0017115, R01 HG003224 and U01 HL81341 to F.P.R.).

Conflict of Interest: none declared.

REFERENCES


