Dependence of Molecular Properties on Proteomic Family for Marketed Oral Drugs

Michal Vieth[§] and Jeffrey J. Sutherland*,[§]

Lilly Research Laboratories and Discovery Informatics, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, Indiana 46285

Received March 31, 2006

Abstract: An association of drugs with their proteomic family reveals that molecular properties of drugs targeting proteases, lipid and peptide G-protein-coupled receptors (GPCRs), and nuclear hormone receptors significantly exceed limits for some properties in the "rule of five", while drugs targeting cytochrome P450s, biogenic amine GPCRs, and transporters have significantly lower values for certain properties. Also, the variation in drug targets appears to be a factor explaining increasing molecular weight over time.

The assessment of "druglikeness" in pharmaceutical research is an important foundation of lead generation and lead optimization programs. Lipinski's "rule of five" established guidelines for the development of permeable compounds.[1,2] These rules and their derivatives[3−5] are routinely used in the selection of compounds for screening. Variation in molecular properties over time has been discussed in the context of increasing attrition from clinical trials,[2,6] and their dependence on route of administration has been highlighted.[7] The importance of the field in drug research is underscored by a continuous stream of review articles.[8−11]

When a compound satisfies the rule of five or one of its derivatives, it has molecular properties similar to those of typical bioavailable drugs. Treated as an aggregate, the pool of oral drugs represents, in an average way, the properties of all biological systems with which they interact, including the drug target. Deviations from the rule of five for certain drug classes have been noted by its authors (i.e., antibiotics, antifungals, vitamins, and cardiac glycosides).[1,2] Since interest in drug targets varies over time, and some targets are pursued more intensely than others, it is reasonable to question whether rules for druglikeness are applicable across different target classes.

We have expanded our marketed, published drug data set to include 33 drugs approved by the FDA after 2003, resulting in a total set of 1756, of which 1210 were administered orally.[7] Only the oral set is considered in this analysis. Six-hundred-forty-two drugs were assigned a gene target and proteomic family. Drugs were assigned to their primary therapeutic target only, even though certain drugs have significant affinity for multiple receptors (e.g., all oral drugs are substrates of metabolizing enzymes, and certain central nervous system drugs act at multiple receptors). Proteomic families group together proteins that share similar functional and structural properties (e.g., kinases, nuclear hormone receptors).[12,13] Targets belonging to the same proteomic family tend to be inhibited by compounds having similar properties,[14] and grouping drugs by family is necessary to understand the characteristics of drugs for most gene targets. We use the assignment of genes to proteome families provided in the Proteome BioKnowledge Library.[15] The families in this analysis include cytochrome P450 receptors (CYP450), G-protein-coupled receptors (GPCR), ion channels, kinases, nuclear hormone receptors (NHR), phosphodiesterases (PDE), proteases, transporters, and others (i.e., those for which a primary target was assigned but not belonging to one of the above proteomic families). We further divide GPCRs into biogenic amine (GPCR-bio), peptide (GPCR-pep), and lipid (GPCR-lipid) varieties, based on the nature of the endogenous ligand.

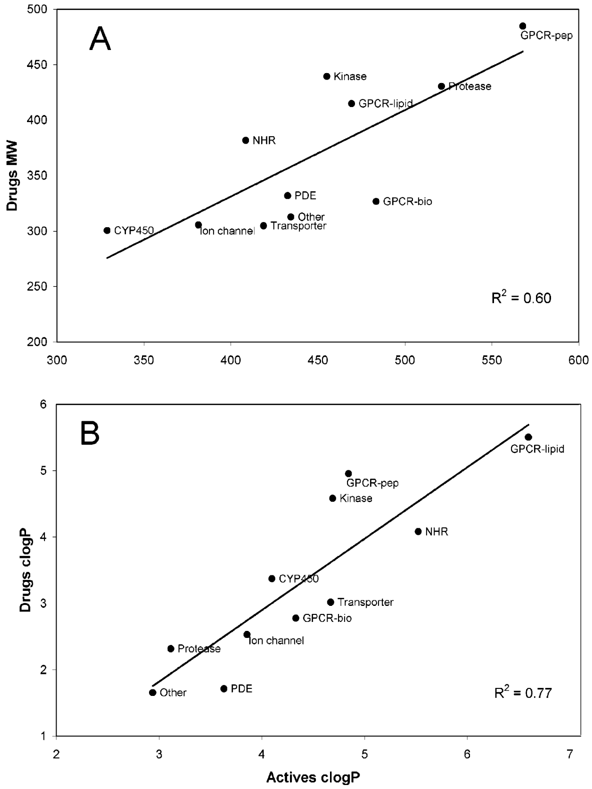

Figure 1. Comparison of the mean of molecular properties (A, MW; B, clogP) between drugs and highly active compounds having K_i or $IC_{50} \leq 10$ nM. The R^2 values for MW, clogP, NHOH, and ONs were 0.60, 0.77, 0.68, and 0.81 respectively. Drugs and actives are categorized according to proteomic family. The number of drugs and actives for each family are given in Table 2.

Two approaches were employed for assigning a primary proteomic family and gene target to as many drugs as possible, utilizing internal curation and the publicly available DrugBank database.[16] The two approaches yield the same family for 90% of 425 (oral and nonoral) drugs having an assignment available from both sources, which rises to 94% when allowing mismatches between GPCRs, transporters, and ion channels (i.e., central nervous system related targets for which multitarget activity occurs frequently) or 89% and 93% over 333 oral drugs only. The combined set of 642 drug-family pairs consists of 508 assignments from our internal curation and 134 from DrugBank. The results of this analysis remain similar whether or not the DrugBank drug-family pairs are included.

The 642 drugs with assigned proteomic family have an average molecular weight (MW) of 335 and calculated log *P* (clogP) of 2.7, compared to 345 and 2.3 for all oral drugs. Averages and 90 percentiles (i.e., the value of a property not exceeded by 90% of drugs) of molecular properties for each family are summarized in Table 1; we focus on 90 percentiiles, from which the rule of five and others were derived. For certain families, it is apparent that molecular properties differ substan-

* To whom correspondence should be addressed. Phone: (317) 655-0833. Fax: (317) 276-6545. E-mail: sutherlandje@lilly.com.

[§] Both authors contributed equally to this work.

10.1021/jm0603825 CCC: $33.50 © 2006 American Chemical Society
Published on Web 05/13/2006

Table 1. Average and 90 Percentiles of Molecular Properties for Oral Drugs Categorized According to Proteomic Family[a]

family (no. targets)	no. drugs	no. (%) passing all 4 RO5	no. (%) passing 3 RO5s	mean MW (*P* value)	mean clogP (*P* value)	mean NHOH (*P* value)	mean ON (*P* value)	90% MW	90% clogP	90% NHOH	90% ON
CYP450 (3)	12	9 (75)	12 (100)	300.5 (0.968)	3.4 (0.858)	0.7 (0.301)	2.9 (0.062)	399.4	8.8	2	5
GPCR-bio (23)	216	188 (87)	213 (99)	326.8 (0.548)	2.8 (0.167)	***1.3 (0.006)***	***4.2 (0.000)***	435.4	5.1	3	7
GPCR-lipid (6)	8	3 (38)	6 (75)	414.9 (0.831)	*5.5 (0.004)*	1.8 (1.000)	5.0 (1.000)	586.2	8.5	3	9
GPCR-pep (3)	11	3 (27)	7 (64)	*484.8 (0.006)*	*5.0 (0.007)*	1.6 (1.000)	*8.5 (0.024)*	600.2	7.5	2	12
ion channel (16)	115	103 (90)	113 (98)	***305.5 (0.029)***	2.5 (0.998)	1.3 (0.058)	4.9 (0.553)	443.2	5.0	2	9
kinase (7)	5	3 (60)	5 (100)	439.4 (0.762)	4.6 (0.423)	2 (1.000)	7.0 (0.982)	493.6	5.6	3	8
NHR (18)	58	38 (66)	58 (100)	381.8 (0.386)	*4.1 (0.000)*	1.4 (0.657)	***3.8 (0.001)***	445.8	7.2	3	6
other (43)	133	114 (86)	126 (95)	312.8 (0.094)	*1.7 (0.042)*	1.9 (0.964)	5.6 (1.000)	448.6	4.4	4	9
PDE (6)	15	15 (100)	15 (100)	331.9 (1.000)	1.7 (0.994)	0.9 (0.556)	6.9 (0.691)	480.2	4.2	2	10
protease (8)	35	24 (69)	29 (83)	*430.6 (0.002)*	2.3 (1.000)	4.5 (0.076)	*7.2 (0.016)*	636.6	5.9	5	11
transporter (8)	37	30 (81)	36 (97)	304.7 (0.581)	3.0 (0.712)	1.3 (0.744)	4.2 (0.141)	423.5	5.5	3	7
all oral	1210	972 (80)	1121 (93)	345	2.3	1.8	5.5	478.4	5.3	4	9

[a] The number of protein targets in each family is indicated in parentheses in column 1. Dunnett's test *P* values with all oral drugs as control are in parentheses. Significant ($P < 0.05$) values are in italic font, with bold text indicating lower than control group values. Dunnett's test *P* values less than 0.05 indicate significant differences in the mean properties for a family with respect to all oral drugs. By use of literature data, in-house target databases, and commercially available drug databases, our curation efforts resulted in the assignment of families to 508 marketed oral drugs. In addition, 901 of 1063 drugs in the publicly available DrugBank database[16] were mapped to our marketed drug database using Smiles strings (543 matches), Chemical Abstract Service (CAS) identifiers (320 matches), and generic names (38 matches), with the additional criteria of molecular weight matches for the last two. SwissProt and GenBank accession numbers from DrugBank were mapped to EntrezGene identifiers using the SRS package (Lion Biosciences, Inc), which were in turn used for assigning a proteomic family using the Proteome database.

Table 2. Comparison of Mean Differences between Actives and Drugs Categorized According to Proteomic Family[a]

family	no. drugs	no. actives less than 10 nM	ΔMW actives − drugs	ΔclogP actives − drugs	ΔNHOH actives − drugs	ΔON actives − drugs
CYP450	12	211	28.3 (0.3056)	−0.7 (0.0933)	0.1 (0.7254)	−0.1 (0.851)
GPCR-bio	216	18101	*156.7 (0.0000)*	*1.6 (0.000)*	*0.6 (0.0338)*	*2.6 (0.0000)*
GPCR-lipid	8	733	54.4 (0.0936)	1.1 (0.0862)	*−0.7 (0.0486)*	−0.4 (0.4873)
GPCR-pep	11	2295	83.0 (0.0879)	−0.1 (0.8397)	0.3 (0.6811)	0.8 (0.5119)
ion channel	115	630	*75.8 (0.0000)*	*1.3 (0.0000)*	−0.1 (0.2445)	0.3 (0.105)
kinase	5	8628	15.7 (0.7963)	0.1 (0.8826)	0.0 (0.9795)	0.1 (0.9257)
NHR	58	2010	*26.7 (0.0209)*	*1.4 (0.0000)*	*−0.2 (0.0107)*	*0.9 (0.006)*
other	133	4887	*121.7 (0.0000)*	*1.3 (0.0000)*	0.3 (0.1381)	*1.4 (0.0001)*
PDE	15	326	*100.7 (0.0000)*	*1.9 (0.0000)*	0.4 (0.099)	0.6 (0.3453)
protease	35	10460	*90.4 (0.0077)*	*0.8 (0.0244)*	−1.5 (0.4084)	*1.9 (0.0332)*
transporter	37	538	*114.0 (0.0000)*	*1.7 (0.0000)*	0.1 (0.5533)	0.6 (0.0985)

[a] Two-tailed *T*-test *P* values are given in parentheses. Significant ($P < 0.05$) values are in italic font.

tially and significantly (*P* values from Dunnett's test less than 0.05) from those of all oral drugs. Drugs in the protease and GPCR-pep families are characterized by significantly higher molecular weight, while those in the ion channel family have lower molecular weight. Drugs in the GPCR-lipid, GPCR-pep, and NHR families have significantly higher clogP. Drugs in the GPCR-pep and protease families have more acceptors, while those in GPCR-bio and NHR have fewer acceptors. Drugs in the GPCR-bio family have significantly fewer donors. In only four families (CYP450, kinase, PDE, and transporter) are the mean values of all four properties statistically similar to those of all oral drugs. Similar observations can be made while looking at the percentage of drugs in each family passing all four (or three of four) original Lipinski rules. GPCR-pep, GPCR-lipid, and protease family targeted drugs have the lowest rule of five compliance (Table 1).

Because of the small number of drugs for certain families, one may question whether the trends noted above would persist for larger samples. The molecular properties of highly active compounds (i.e., K_i or $IC_{50} \leq 10$ nM) from literature sources (as curated by GVK Biosciences, www.gvkbio.com) have been compared to those of drugs (Table 2, Figure 1). While the majority of actives have higher values of properties than drugs, the trends among families observed for drugs are generally observed for highly active compounds. Among the GPCR-pep, kinase, and CYP450 actives, all four RO5 properties are not statistically different from corresponding drugs, while in the GPCR-bio and NHR family actives, all four properties are significantly higher than drugs. Differences between highly active compounds and drugs have been noted in the literature.[17]

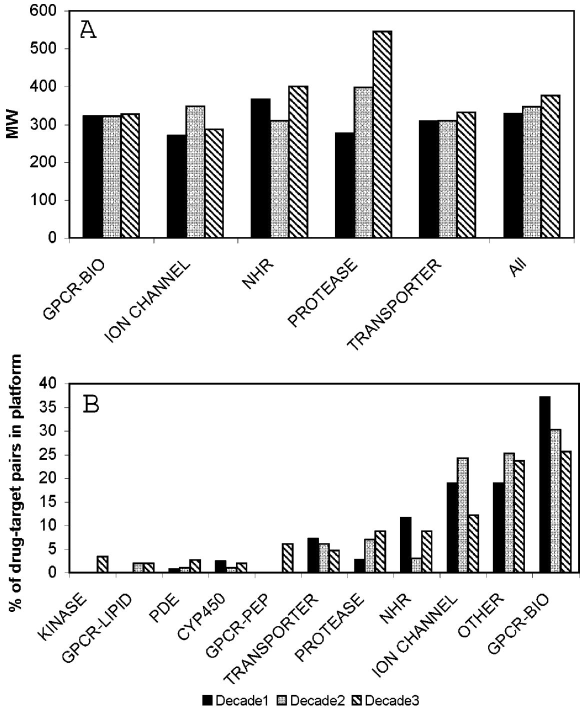

Figure 2. (A) Average molecular weight for proteomic families having more than one drug for each of 3 decades. The number of drugs in each decade are GPCR-bio (92-30-38), ion channel (47-24-18), NHR (29-3-13), protease (7-7-13), transporter (18-6-7), all (389-136-197). (B) Percentage of oral drugs targeting receptors in each family for each of 3 decades before 1982 (black bars), 1982−1992 (gray bars), and 1993−2005 (hashed bars).

These differences may arise from a lack of bioavailability/permeability for actives, as previously suggested;[18] activity in an in vitro or functional assay does not require the compound to be bioavailable.

Where the trends among drugs differ significantly from trends among highly active compounds, the selection of family-tailored ranges is open to debate. We prefer ranges deduced from drugs, especially with regard to the GPCR-bio and NHR families, both of which are represented by a substantial number of drugs. General similarities in trends between drugs and highly active compounds suggest that the variations in molecular properties for drugs arise from the properties of the binding pockets of the targeted receptors (e.g. NHRs have lipophilic pockets,[19] and their ligands have higher lipophilicity, as this analysis shows).

A number of researchers have highlighted an upward trend in MW and lipophilicity between older and newer drugs.[2,6] This trend has been discussed in the context of higher attrition from clinical trials.[6] Of the 642 drug-family pairs in this analysis, 495 have been annotated with the decade of FDA approval.[20] We investigated the variation of MW over time for families having more than one example of drugs in each of 3 decades/time periods (before 1982, 1982−1992, 1993−2005) (Figure 2A). With recognition of the very small sample sizes for many family-decade combinations, it appears that only drugs targeting the protease family have increased substantially in MW with passing decades. Rather, we suggest that changes in MW over time result from variations in the target portfolios of pharmaceutical companies (Figure 2B). Most notably, a significant decrease of biogenic amine GPCR drugs in the recent decades (43−28%) and increases in protease and peptidic GPCR targeted drugs may explain much of the overall MW trend. Variation in properties over time for a given family may result from varying pharmaceutical interest in its members (e.g., serine proteases, metalloproteases, etc.).

The central assumption in the applicability of standard rules for druglikeness in a screening program is that the target of interest requires molecular properties similar to those of the average drug. Since bioavailability results from the interactions of drugs with the same biological systems (e.g., those that collectively determine ADME properties), it is plausible that well-defined ranges of molecular properties can account for favorable interactions with those systems. For certain proteomic families, application of standard rules of druglikeness would bias screening collections away from the required molecular properties for achieving high affinity. The need to balance bioavailability and affinity suggests that modified rules of druglikeness be adopted for certain target classes.

Acknowledgment. We thank Richard E. Higgs for consultation on the statistical data analsyis.

Supporting Information Available: Tabulated molecular properties for 1210 drugs in an Excel file, with family assignments and accession numbers for public databases. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(2) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46* (1−3), 3−26.

(3) Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000**, *14* (3), 251−264.

(4) Kelder, J.; Groothenhuis, P. D. J.; Bayada, D. M.; Delbressine, L. P.; Ploemen, J.-P., Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Res.* **1999**, *16*, 1514−1519.

(5) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.* **1999**, *1* (1), 55−68.

(6) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A comparison of physiochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* **2003**, *46* (7), 1250−1256.

(7) Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; Savin, K. A.; Durst, G. L.; Hipskind, P. A. Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.* **2004**, *47* (1), 224−232.

(8) Clark, D. E.; Pickett, S. D. Computational methods for the prediction of "drug-likeness". *Drug Discovery Today* **2000**, *5* (2), 49−58.

(9) Egan, W. J.; Walters, W. P.; Murcko, M. A., Guiding molecules towards drug-likeness. *Curr. Opin. Drug Discovery Dev.* **2002**, *5* (4), 540−549.

(10) Lajiness, M. S.; Vieth, M.; Erickson, J. Molecular properties that influence oral drug-like behavior. *Curr. Opin. Drug Discovery Dev.* **2004**, *7* (4), 470−477.

(11) Muegge, I. Selection criteria for drug-like compounds. *Med. Res. Rev.* **2003**, *23* (3), 302−321.

(12) Mirzabekov, A.; Kolchinsky, A. Emerging array-based technologies in proteomics. *Curr. Opin. Chem. Biol.* **2002**, *6* (1), 70−75.

(13) Swindells, M. B.; Overington, J. P. Prioritizing the proteome: identifying pharmaceutically relevant targets. *Drug Discovery Today* **2002**, *7* (9), 516−521.

(14) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biospectra analysis: model proteome characterizations for linking molecular structure and biological response. *J. Med. Chem.* **2005**, *48* (22), 6918−6925.

(15) Hodges, P. E.; Carrico, P. M.; Hogan, J. D.; O'Neill, K. E.; Owen, J. J.; Mangan, M.; Davis, B. P.; Brooks, J. E.; Garrels, J. I. Annotating the human proteome: the Human Proteome Survey Database (HumanPSD) and an in-depth target database for G protein-coupled receptors (GPCR-PD) from Incyte Genomics. *Nucleic Acids Res.* **2002**, *30* (1), 137−141.

(16) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668−D672 (database issue)

(17) Oprea, T. I. Current trends in lead discovery: are we looking for the appropriate properties? *J. Comput.-Aided Mol. Des.* **2002**, *16* (5−6), 325−334.

(18) Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432* (7019), 855−861.

(19) Wurtz, J. M.; Bourguet, W.; Renaud, J. P.; Vivat, V.; Chambon, P.; Moras, D.; Gronemeyer, H. A canonical structure for the ligand-binding domain of nuclear receptors. *Nat. Struct. Biol.* **1996**, *3* (1), 87−94.

(20) Food and Drug Administration Orange Book. http://www.fda.gov/cder/orange/obreadme.htm.

JM0603825