OPINION

The druggable genome

Andrew L. Hopkins and Colin R. Groom

An assessment of the number of molecular targets that represent an opportunity for therapeutic intervention is crucial to the development of post-genomic research strategies within the pharmaceutical industry. Now that we know the size of the human genome, it is interesting to consider just how many molecular targets this opportunity represents. We start from the position that we understand the properties that are required for a good drug, and therefore must be able to understand what makes a good drug target.

Biological systems contain only four types of macromolecule with which we can interfere using small-molecule therapeutic agents: proteins, polysaccharides, lipids and nucleic acids. Toxicity, specificity and the inability to obtain potent compounds against the latter three types means that the vast majority of successful drugs achieve their activity by binding to, and modifying the activity of, a protein. This limits the molecular targets for which commercially viable compounds can be developed, leading to the concept of 'the druggable genome' — the subset of the ~30,000 genes in the human genome that express proteins able to bind drug-like molecules.

One way of assessing the opportunities available to the pharmaceutical industry is to begin by studying the properties that are required in a commercially viable drug. For the most part, this means an orally bioavailable compound. The physico-chemical properties that are necessary to increase the likelihood of oral bioavailability have been formalized into the 'rule-of-five'[1] (BOX 1). Constraints such as these dictate the type of protein we see as drug targets — simply put, drug targets need to be able to bind compounds with appropriate properties.

Druggable protein families

The druggable subset of the human genome can be predicted using several methods. In a comprehensive review of the accumulated portfolio of the pharmaceutical industry, Drews[2,3] identified 483 targets, and concluded that there could be 5,000–10,000 potential targets on the basis of an estimate of the number of disease-related genes[4]. However, this analysis did not focus on the properties of the drugs that define those targets. The idea of assessing the number of ligand-binding domains has also recently been introduced as a measure of the number of potential points at which small-molecule therapeutic agents could act — suggestions are that this figure could be even greater than 10,000 (REF. 5).

Binding sites on proteins usually exist out of functional necessity; therefore, most successful drugs achieve their activity by competing for a binding site on a protein with an endogenous small molecule. For a drug to be effective, it must bind to its molecular target with a reasonable degree of potency. Our analysis of the Investigational Drugs Database (produced by Current Drugs) and the Pharmaprojects Database (produced by PJB Publications), in addition to a thorough review of the literature, identifies 399 non-redundant molecular targets that have been shown to bind rule-of-five-compliant compounds with binding affinities below 10 μM.

Although there is some degree of overlap with earlier work[2–4], we have captured several proteins that are targeted by experimental drugs, and eliminated some targets for which activity has not yet been shown to be modulated by rule-of-five-compliant compounds. Most of the drugs and leads that were identified in this survey are competitive with an endogenous ligand at a structurally defined binding site.

We have taken the sequences of the drug-binding domains of these proteins and determined the families that they represent, as captured by their InterPro domain[6,7]. Only 130 protein families represent the known drug targets (ONLINE TABLE 1). Nearly half of the targets fall into just six gene families: G-protein-coupled receptors (GPCRs), serine/threonine and tyrosine protein kinases, zinc metallopeptidases, serine proteases, nuclear hormone receptors and phosphodiesterases (FIG. 1a).

Box 1 | **Guidelines for oral bioavailability: the 'rule-of-five'**

The 'rule-of-five' analysis by Lipinski *et al.*[1] shows that poor absorption or permeation of a compound are more likely when: there are more than five hydrogen-bond donors; the molecular mass is more than 500 Da; the lipophilicity is high (expressed as cLogP > 5); and the sum of nitrogen and oxygen atoms is more than 10. These rules, more appropriately described as guidelines, do not cover drugs that are derived from natural products, for which other absorption mechanisms are involved.

Clearly, published data on the oral bioavailability of existing drugs could be used as a method for defining the properties of viable drugs; however, our approach using the rule-of-five allows predictions to be made. In practice, the number of targets identified by applying the rule-of-five filters differs little from that obtained solely by literature analysis of all known drugs, whether rule-of-five compliant or not.

Table 1 | **Comparison of the druggable genomes of selected eukaryotes**

	Homo sapiens	*Drosophila melanogaster*	*Caenorhabditis elegans*	*Saccharomyces cerevisiae*
Total number of predicted genes[8,9,16]	~30,000	13,601	18,424	6,241
Number of proteins in proteome*	21,688	13,849	17,946	6,127
Number of estimated druggable targets	3,051	1,714	2,267	508
Percentage that are predicted druggable targets	~10–14%	12%	12%	8%

Three hundred and seventy-six targets identified to bind rule-of-five-compliant drugs have had InterPro domains assigned. The prevalence of these InterPro domains in various genomes has then been determined. Twenty-three more bacterial and viral drug targets for which InterPro assignments could not be made have not been included in any of our analyses. *Data taken from InterPro, 29 October 2001.

The sequence and functional similarities within a gene family are usually indicative of a general conservation of binding-site architecture between family members. This would suggest that if one member of a gene family were able to bind a drug, other members would also be able to bind a compound with similar physico-chemical properties. Using this reasoning, 3,051 of the predicted 30,000 or so genes in the human genome[8,9] code for a protein with some precedent for binding a drug-like molecule (FIG. 1b, ONLINE TABLE 1). A comparative analysis of the eukaryotic genomes of worm, fly and yeast also reveals that approximately one in ten of the proteins expressed by these genomes belong to gene families with members that have previously shown modulation by small-molecule drugs (TABLE 1).

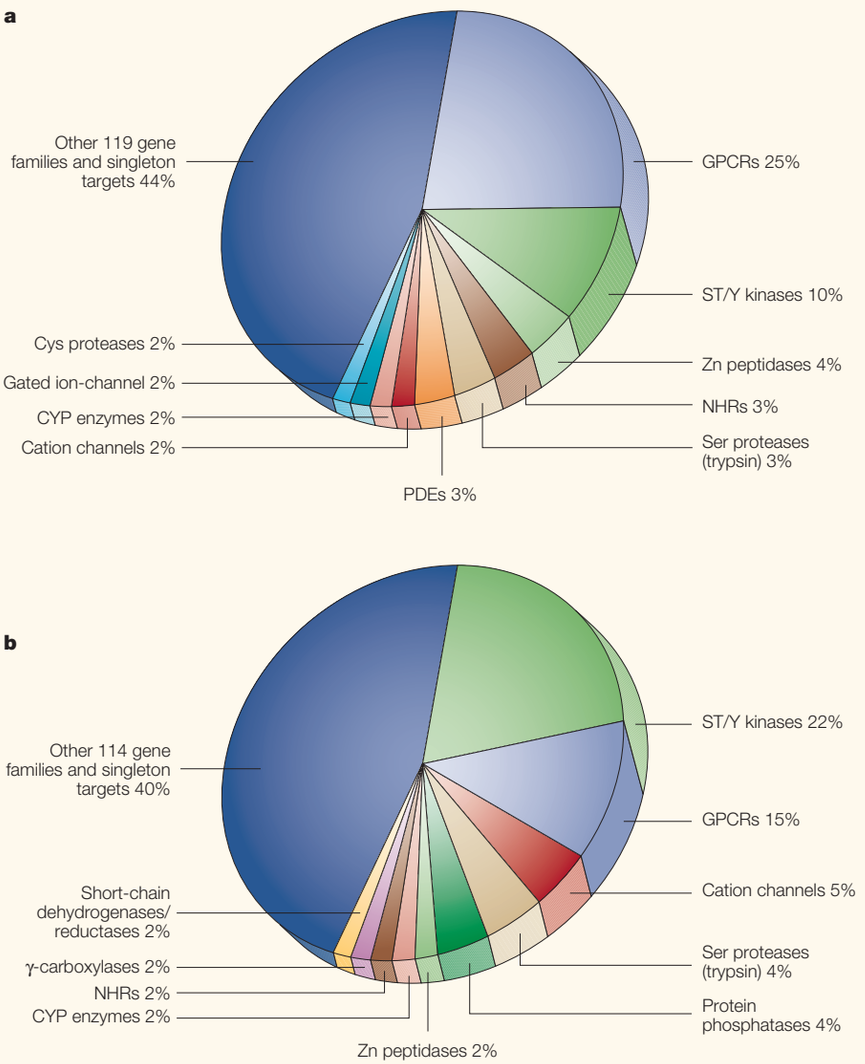

Figure 1 | **Drug-target families.** Gene-family distribution of **a** | the molecular targets of current rule-of-five-compliant experimental and marketed drugs, and **b** | the druggable genome. Serine (Ser)/threonine and tyrosine protein kinases are grouped as one gene family (ST/Y kinases), as are class 1 and class 2 G-protein-coupled receptors (GPCRs). CYP, cytochrome P450; Cys, cysteine; NHR, nuclear hormone receptor; PDE, phosphodiesterase; Zn, zinc.

Expanding the druggable genome

At present, approximately half of the proteins expressed by the genome are functionally unclassified, and of course, some of these might prove to be druggable. However, it is clear from the distribution of the gene-family populations that there are no undiscovered large protein families, which indicates that remaining targets will be members of very small families. Clearly, the number of potential protein targets could be larger than the number of genes, owing to post-translational modifications and assembly of functional complexes; however, this is not likely to increase the number of specific drug-binding sites.

Further evidence for this can be drawn from the observation that despite numerous screening attempts, many targets have failed to show any evidence of binding compounds that are potent and 'drug-like'. This might be a function of the chemical diversity of corporate compound files. However, if druggability is an inherent property of the protein, then *a priori* assessment criteria of potential targets to assess the likelihood of developing a drug against a particular site can be developed. As most drugs bind to discrete binding sites, which can be identified readily by structural analysis, it is possible to filter what we term 'beautiful binding sites' from the wealth of protein-structure data that are available in the Protein Data Bank and are expected soon from structural genomics projects.

As most drugs compete against small molecules for binding sites on proteins, the number of these binding sites is probably a function of the size of the metabolome (the total set of small molecules in an organism). One route to target discovery might therefore

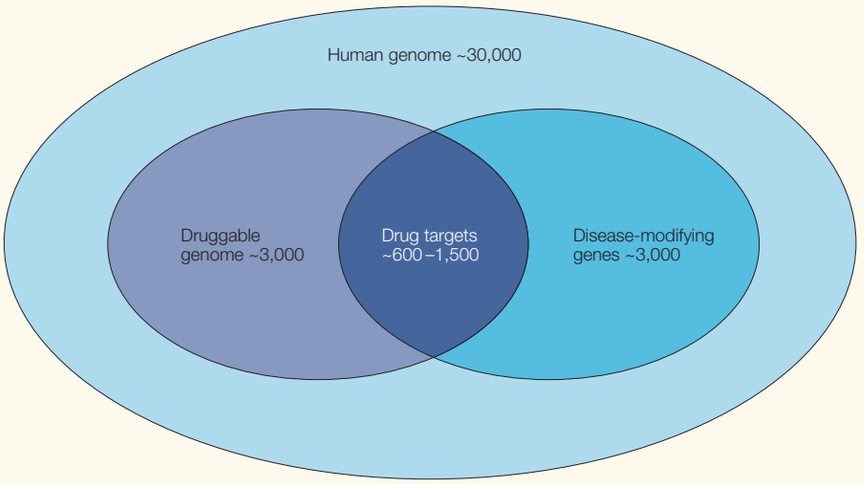

Figure 2 | **Number of drug targets.** The effective number of exploitable drug targets can be determined by the intersection of the number of genes linked to disease and the 'druggable' subset of the human genome.

lie in identifying enzymes and receptors from metabolomic profiling[10–12]. By contrast, the druggability of targets identified by proteomic or transcription-profiling studies is likely to be low.

Druggable does not equal drug target

The ability of a protein to bind a small molecule with the appropriate chemical properties at the required binding affinity might make it druggable, but does not necessarily make it a potential drug target, for that honour belongs only to proteins that are also linked to disease.

Recent estimates propose that there are from 3,000 (REF. 13) to 10,000 (REF. 4) disease-related genes, and large-scale mouse-knockout studies have revealed that only ~10% of all gene knockouts might have the potential to be disease modifying[14], which supports estimates at the lower end of this range.

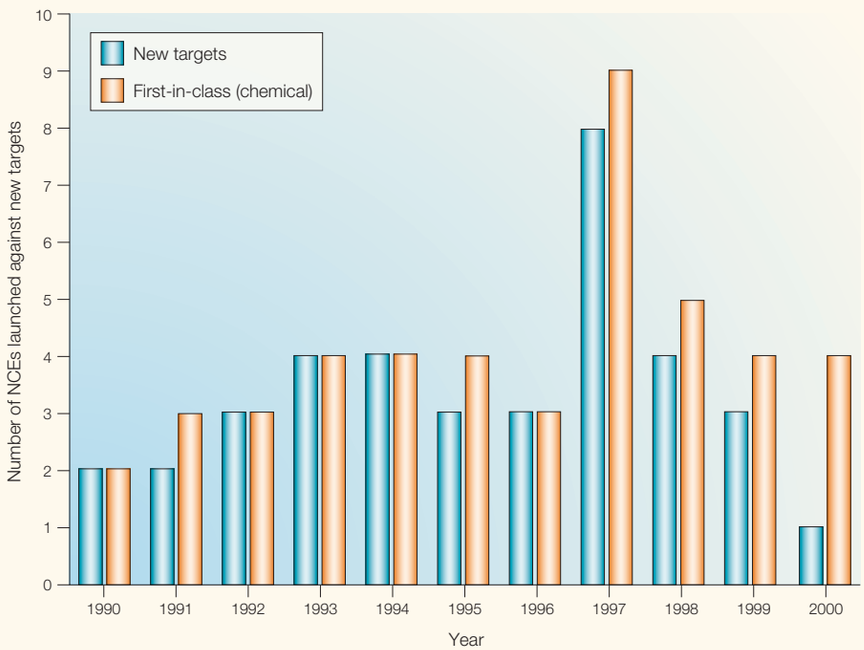

Figure 3 | **Novel drug launches.** The graph shows the number of small-molecule, 'first-in-class' drugs and associated new drug targets that have been launched on the market in the past decade (data derived from collating annual 'This Year's Drugs' reviews of *Drug News & Perspectives,* Prous Science). NCE, new chemical entity.

The potential drug targets that the pharmaceutical industry can exploit are captured in the intersection between the druggable genome and those genes related to disease, as shown in FIG. 2. An analysis of the antifungal targets from the yeast genome indicates that the intersection might be as small as 2–5% of the genome (C.R.G. and J. E. Mills, unpublished observations) — extrapolating to man, this suggests a total of 600–1,500 small-molecule drug targets.

Targets to market

Despite the massive increases in research and development (R&D) investment over the past decade, and the advent of molecular biology, the rate at which drug targets are clinically validated and brought to market is growing rather slowly. On average, new drugs are launched against only four novel targets each year (FIG. 3).

The distribution of target types shown in FIG. 1a is similar to the distribution seen in the original work of Drews[2,3], but to our surprise, of our set of 399 targets with known rule-of-five-compliant agents, we could identify only 120 proteins as the targets of drugs that are actually marketed. This small number of targets calls into question the common assumption that a large number of targets are necessary to build a successful industry[13]. Differentiation between drugs that bind to the same receptor could lead to the development of several distinct classes, targeting a range of diseases.

The overall distribution of launched targets by biochemical class is similar to that observed for all targets with drug-like leads (FIG. 4). Enzymes represent just under half of the launched targets (47%), whereas GPCRs account for 30%. All other classes, such as ion channels and nuclear hormone receptors, account for less than a quarter of the identified launched targets.

Implications

Commercial pressure forces the pharmaceutical industry to focus on developing orally bioavailable small molecules, limiting opportunity to the number of binding sites for such molecules on proteins encoded by the genome. New mechanisms, such as protein drugs, antibody therapies, DNA vaccines and non-oral drug delivery systems, could expand the range of potential targets to those fundamentally not tractable with rule-of-five-compliant therapies.

A comparison of gene-family size with the number of targets in a family that have specific leads shows that many large, druggable gene families are still under-exploited (FIG. 5). The application of high-throughput screening

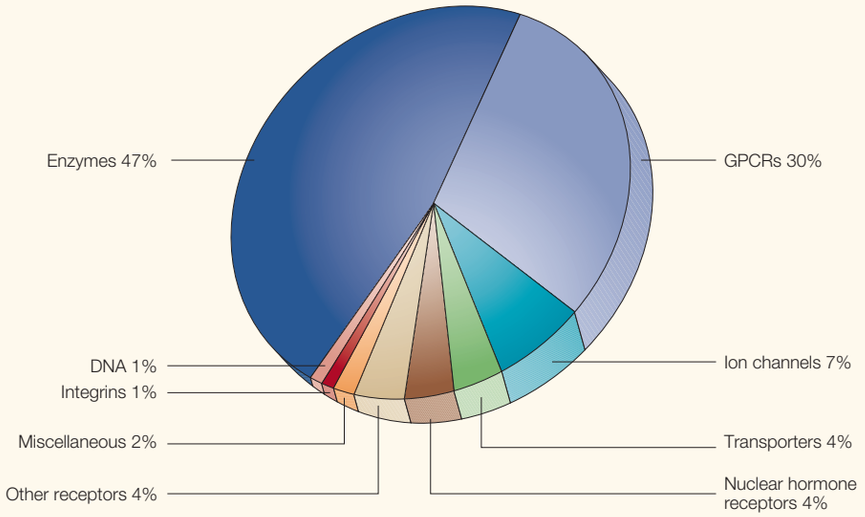

Figure 4 | **Marketed small-molecule drug targets by biochemical class.** GPCR, G-protein-coupled receptor.

in the pharmaceutical industry and the limited number of druggable targets suggest that, within the next decade, the industry could reach a position in which 'hits' or chemical leads are available for most potentially druggable targets. The challenge for the industry will then not necessarily be in the discovery of leads, but in discovering and assessing the therapeutic utility of its leads and druggable targets.

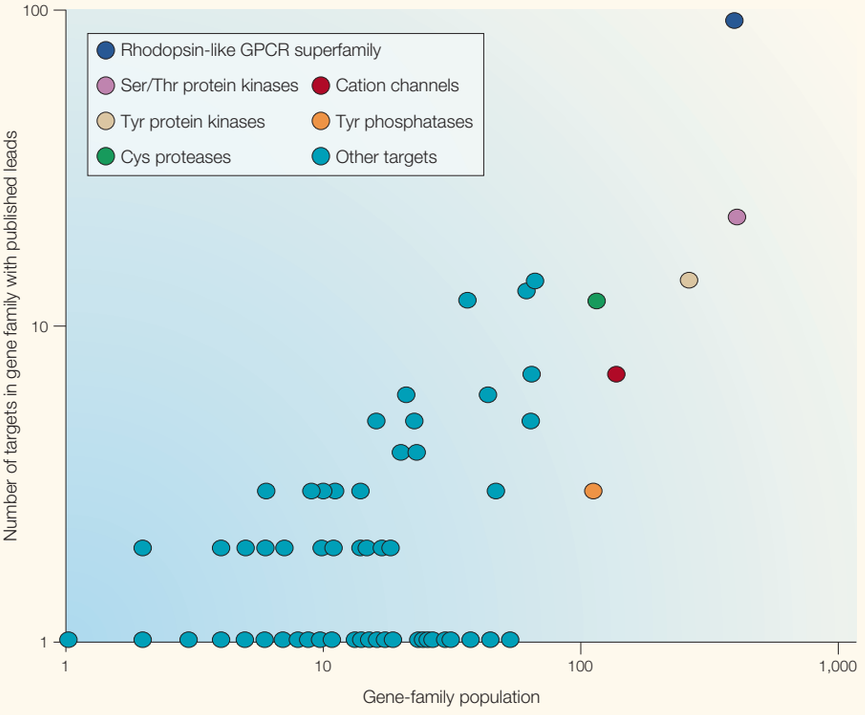

Figure 5 | **Exploitation of the genome, as measured by known leads.** In the past few decades, the pharmaceutical industry has assigned high priority to research into gene families, such as kinases, matrix metalloproteinases and cysteine proteases. Few drugs aimed at these gene families have yet reached the market, although many are progressing through development. Cys, cysteine; GPCR, G-protein-coupled receptor; Ser, serine; Thr, threonine; Tyr, tyrosine.

The limited number of small-molecule drug targets suggests that to exploit the opportunity of the druggable genome in a cost-effective manner, the next round of innovation for the pharmaceutical industry lies not necessarily just in the science, but also in the business models[15].

Andrew L. Hopkins and Colin R. Groom are at the Molecular Informatics, Structure and Design Department, Pfizer Global Research & Development, Sandwich, Kent, CT13 9NJ, UK.
Correspondence to C.R.G.
e-mail: colin_r_groom@sandwich.pfizer.com

doi: 10.1038/nrd892

1. Lipinski, C., Lombardo, F., Dominy, B. & Feeney, P. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**, 3–25 (1997).
2. Drews, J. Genomic sciences and the medicine of tomorrow. *Nature Biotechnol.* **14**, 1516–1518 (1996).
3. Drews, J. & Ryser, S. Classic drug targets. *Nature Biotechnol.* **15**, 1318–1319 (1997).
4. Drews, J. Drug discovery: a historical perspective. *Science* **287**, 1960–1964 (2000).
5. Bailey, D., Zanders, E. & Dean, P. The end of the beginning for genomic medicine. *Nature Biotechnol.* **19**, 207–209 (2001).
6. Apweiler, R. *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29**, 37–40 (2001).
7. *Proteome Analysis Database* [online], (version analysed with release date 29 Oct 01) <http://www.ebi.ac.uk/proteome/> (2001).
8. Lander, E. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
9. Venter, J. C. *et al*. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
10. Tweeddale, H., Notley-McRobb, L. & Ferenci, T. Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ('metabolome') analysis. *J. Bacteriol.* **180**, 5109–5119 (1998).
11. Fiehn, O. *et al.* Metabolite profiling for plant functional genomics. *Nature Biotechnol.* **18**, 1157–1161 (2000).
12. Raamsdonk, L. *et al.* A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnol.* **19**, 45–50 (2001).
13. Claverie, J.-M. What if there are only 30,000 human genes? *Science* **291**, 1255–1257 (2001).
14. Walke, D. W. *et al. In vivo* drug target discovery: identifying the best targets from the genome. *Curr. Opin. Biotechnol.* **12**, 626–631 (2001)
15. Lehman Brothers. *The Fruits of Genomics* (Lehman Brothers, 2001).
16. Rubin, G. Comparative genomics of the eukaryotes. *Science* **24**, 2204–2215 (2000).

Acknowledgements
We are indebted to J. P. Overington (Inpharmatica, London), A. Alex and L. Beeley for their contributions to the ideas on the physico-chemical limits for protein-binding sites. We also thank R. W. Spencer and C. Lipinski (Pfizer, Groton, Connecticut, USA) for much stimulating discussion.

Online links

FURTHER INFORMATION
InterPro: http://www.ebi.ac.uk/interpro/search.html
Protein Data Bank: http://www.rcsb.org/pdb/
Proteome Analysis Database:
http://www.ebi.ac.uk/proteome/
Access to this interactive links box is free online.