feature

The druggable genome: an update

Andreas P. Russ* and Stefan Lampel,
*andreas.russ@bioch.ox.ac.uk

Now that a finished sequence of the human genome with high-quality annotation [1] is available, it is a good time to take a fresh look at the druggable genome. Our analysis suggests that there is a druggable gene count between 2000 and 3000, in general agreement with previous estimates (~3000). However, there is evidence of a significant shift in the contribution from the major target families [there are fewer rhodopsin-like G-protein-coupled receptors (GPCRs) and protein kinases but more proteases than expected].

What is druggability?

The druggability of protein families has been discussed in several excellent articles [2–4] and can be described, in a nutshell, as the presence of protein folds that favour interactions with drug-like chemical compounds [2]. Proteins lacking these structural features might have interesting biological properties but are unlikely to be readily amenable to pharmaceutical modulation. The problem can be visualized as doors and keyholes. A door (protein) might control access to an interesting pathway but if it does not have the appropriate keyhole (druggable domain) it cannot be opened. Analysis of druggability, so to speak, is the analysis of keyholes [5].

This analogy also illustrates that the concept of druggability reflects our current technical abilities [6]. We cannot recognize structures that might be perfect keyholes for a future technology, just as swipe-card readers would not have been considered to be keyholes 50 years ago. However, the number of proteins that are tractable with current technology should be an important aspect of portfolio management.

It is also important to keep in mind the distinction between the druggability of a protein and its actual qualities as a drug target [2,7]. Many proteins are druggable, according to their structure, but modulating their biological function will not provide any therapeutic benefit; not every door with a keyhole leads to a desirable place. The actual drug targets are the subset of druggable proteins that possess structural and functional features of druggability. Their only real validation comes with successful clinical use.

The prediction of structural druggability by protein-sequence annotation is discussed in this article. It must not be confused with structural genomics, the determination of detailed protein structures with high-throughput methods.

How can we define the druggable genome?

A census of druggable proteins is a moving target because tools and resources are evolving rapidly. Over time, the estimates of potential targets have been reduced substantially to reflect the lower than expected number of protein-coding genes and they are now converging at ~3000 druggable loci [2,4,8,9]. The main parameters influencing the count are: the coverage of sequence databases, the tools used for sequence annotation, structural information about tractable folds, bioinformatics tools and biological information about protein function (Box 1).

One crucial parameter, the human genome sequence itself, has now stabilized. Although the first working draft sequence covered only 90% of the human genome, the human genome sequence (build 35) covers 99% of the euchromatic genome with high-quality sequence and there are only 341 gaps remaining [1]. Thus, current estimates should only miss 1% of potential targets because of a lack of database representation.

Genome annotation relies on correct gene predictions to provide hypothetical protein sequences. Databases of cDNAs and expressed sequence tags (ESTs) have the potential to overestimate the gene count because cloning artefacts and sequencing errors can create unique sequence entries that are not biologically relevant. Therefore, a nonredundant set of gene models has to be benchmarked to a nearly complete genome sequence.

Even with a perfect set of predicted proteins the annotation of druggable domains is not unambiguous. Although several excellent algorithms for domain annotation are established, by their very nature they are statistical methods and, depending on the parameters chosen, they will provide conservative estimates with high confidence or more-aggressive, higher counts that enter the 'twilight zone' of sequence homology. Thus, there cannot be one correct count, only estimates at different levels of confidence.

As well as all of the technical limitations, biological complexity has to be taken into account. Multimeric protein complexes and successful, promiscuous drugs (drugs acting on more than one target) do not allow the immediate generalization from the gene to the drug target (Box 2). When discussing the merits of druggable genome annotation, it is very important to keep all of the limitations in mind.

How can we measure druggable space?

An influential paper by Hopkins and Groom [2] defined a set of protein domains used by

BOX 1

Technical limitations of protein prediction and annotation

The algorithms used to assign protein domains by primary sequence data are very sophisticated and specific and protein domain databases, for example Interpro [11] and PFAM [12], are very advanced. The limiting factors in the analysis of protein-coding genes are the quality of the underlying whole genome data and the strategies used to move from the raw genomic sequence to the validated protein sequence. Although it is straightforward to assign known druggable domains to a protein sequence, it is much more challenging to move from genomic data to reliable mRNA and protein-sequence prediction. There is no perfect algorithm for this (yet) and the current best practice is to use complementary computational approaches to give the best possible results [10,18].

Despite the finished human genome sequence, an unambiguous gene count is currently not possible. Genes can be missed because they are not represented in the genome database [1% not covered in the human genome sequence (build 35)], they contain sequencing errors or they are not detected by gene-prediction programs. Overestimates can result from pseudogenes or false-positive annotation by gene-prediction programs.

1359-6446/05/$ – see front matter ©2005 Elsevier Ltd. All rights reserved. PII: S1359-6446(05)03666-4

feature

BOX 2

One gene, one target?

A fundamental problem with estimating the number of druggable targets is how multimeric proteins and protein isoforms are treated. Although most of us, the authors included, would like to think along a 'one gene, one target' model, it is obvious that this can only be a rough approximation. In some target classes (e.g. ion channels) the active biological principle is a heteromeric multiprotein complex composed of subunits encoded by different genes. Although this complex is one well-defined target for the pharmacologist, the genomic annotation can only count the number of genes encoding the known subunits. It is difficult to extrapolate from this count to several pharmacological targets because the same subunits can unite in different combinations to form different targets. Detailed estimates of isoforms of multi-protein complexes are beyond the scope of this article, which provides only estimates of subunit numbers. Similar considerations apply to splice variants and other protein isoforms. In other words, the genomic analysis cannot replace the wet-bench biologist or geneticist or the pharmacologist (in defining the actual biological target) but has to be viewed as a specific form of information technology to support biological and pharmacological studies.

Another aspect of the 'one gene, one target' question is the fact that many successful drugs have more than one molecular target and their therapeutic utility is determined by the balance of their actions and not their absolute specificity [19,20]. The unexpected long-term clinical effects of COX-2-specific inhibitors provide a prominent example illustrating this complexity.

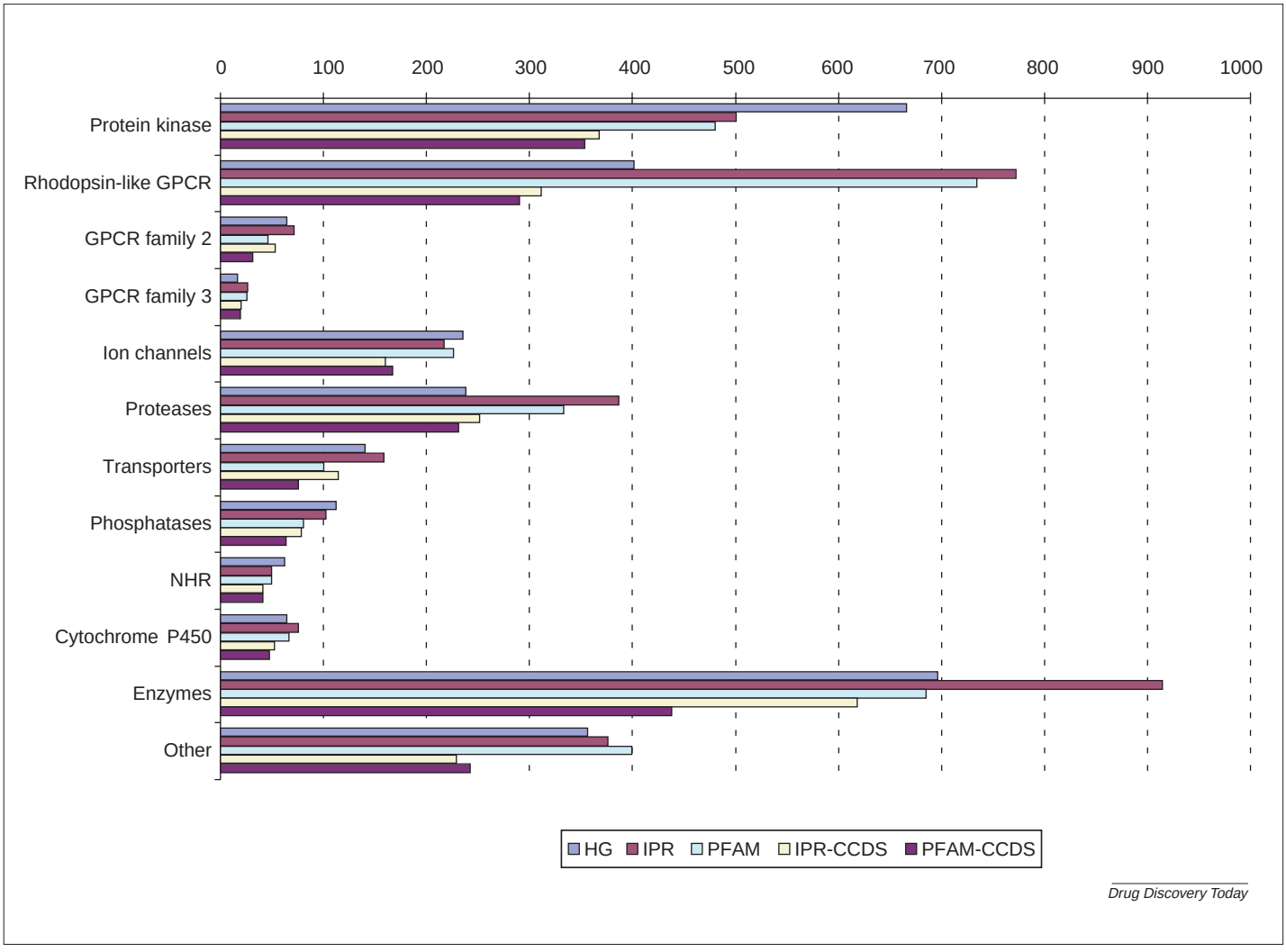

FIGURE 1

The count of druggable protein classes based on Ensembl and CCDS annotations of the human genome sequence (build 35) compared with the data of Hopkins and Groom (HG) [2]. IPR represents the Interpro annotation of the full Ensembl gene set. PFAM represents the PFAM annotation of full Ensembl gene set. IPR-CCDS represents the Interpro annotation of full CCDS gene set. PFAM-CCDS represents the PFAM annotation of full CCDS gene set. The sum of druggable domains exceeds the count of druggable genes because some proteins contain more than one druggable domain. IPR and PFAM predictions of rhodopsin-like GPCR include ~400 sensory receptors. Abbreviation: NHR, nuclear hormone receptors.

feature

established and experimental small-molecule drugs, providing a gene count based on the public and private genomic datasets that were available at the time. We are attempting to provide an update and re-evaluation of this dataset using only publicly available genome sequences and annotations. Details of our search strategy and our results are provided as supplementary information on our webpage (www.bioch.ox.ac.uk/genetics/Russ/russ.html).

Several large-scale efforts for genome annotation are currently under way. Different strategies to implement the annotation pipeline and slightly different parameter sets (e.g. cut-off thresholds for significance) result in nonidentical annotations of the same genome sequence. Our analysis is based on build 35 of the human genome sequence and the automated whole-genome annotation provided by the Ensembl Project [10] as well as the Consensus Chemical Database Service (CCDS) consensus annotation (www.ncbi.nlm.nih.gov/CCDS). The CCDS consensus annotation is the subset of genes that have been consistently predicted by all major genome databases and it can be considered as the more conservative, high-stringency prediction that will probably be an underestimate.

Hopkins and Groom and other authors [4] have used the Interpro (IPRO) classification of protein domains. IPRO [11] is a meta-database that compiles data from different algorithms to classify protein sequences (e.g. PFAM, the protein families database, and PRINTS, a compendium of protein fingerprints). We used an updated version of their IPRO domain set, resulting in a total druggable gene number of 3533 on Ensembl (release 32) and 2225 on CCDS (February 2005).

Upon inspection, data show that most predictions appear to be correct but that the set contains several, significant false-positive predictions (i.e. contains genes that scored positive in the automated annotation because of partial matches but do not contain a complete druggable domain). The most important factor for overprediction is the inclusion of olfactory and taste receptors, which are highly similar to rhodopsin-like GPCRs. Substituting this group with a manually curated set of 290 GPCRs reduces the count to 3050 (from 3533 on Ensembl) and 2204 (from 2225 on CCDS).

To address the problem of overprediction, we performed similar searches using the PFAM protein domain classification [12]. 108 of the IPRO definitions used by Hopkins and Groom [2] are directly based on PFAM models and we manually chose equivalent models for the remaining domains, leading to a set of 182 PFAM domains. This approach resulted in fewer false-positive results, returning a count of 2917 on Ensembl (release 32) and 1942 on CCDS, after correcting for sensory receptors.

Is the power shifting between the ruling families?

At first glance these numbers are close to numbers that have been previously published. So what's new? Breaking the total count down into functional categories reveals that some predictions are, indeed, very stable but others seem to have changed significantly (Figure 1).

The two largest families, protein kinases and rhodopsin-like GPCRs, still top the league table but numbers of them are lower than expected. IPRO and PFAM predictions are very close and estimate the count for kinases at 480–500, down more than 20% from earlier predictions and in close agreement with the detailed annotation of the kinome (518) [13]. Obtaining an exact count for rhodopsin-like GPCR drug targets is cumbersome because hundreds of closely related sensory receptors inflate the automated domain count but will probably not have any therapeutic potential. Our own manual curation identified ~280 nonsensory members of the family, in agreement with the IUPHAR receptor database (http://iuphar-db.org/iuphar-rd/index.html) and other recent publications [14] predicting <300 druggable members of this family. These receptors, many of them linked to disease-relevant physiology, are still most likely to be the richest opportunity for drug discovery in the short- to mid-term but the lower number suggests that the area might soon be saturated by drug discovery efforts.

The counts for non-rhodopsin GPCRs, ion channels, transporters, nuclear hormone receptors and cytochromes are largely unchanged, indicating that these smaller families are annotated in great detail already. The most important group that consistently appears larger than previously expected is the protease group. Even the most stringent of our counts (PFAM–CCDS), probably an underestimate, yields ~230 putative proteins. Applying the IPRO models increases the count to ~380. The higher numbers are consistent with recent curated counts [15], which predict 553 proteases and related proteins (although they include additional domain signatures that are not defined as druggable). It appears that the druggable protease space might be at least as large as that of the rhodopsin-like GPCRs.

The heterogeneous group of druggable enzyme families seems to be of a similar size to that previously predicted but, because of its complexity, a detailed validation of this observation is beyond the scope of this article.

Figure 2 compares previous estimates of the druggable genome with an optimistic and conservative interpretation of the current dataset. We assume that the numbers for kinases, GPCRs and smaller target families will remain stable. In the optimistic scenario, the larger number of proteases compensates for the reduction of other families and arrives at just over 3000 targets, the same total as Hopkins and Groom reported in 2002 [2]. The conservative count uses the validated protease subfamilies and high-stringency predictions for enzymes and other target classes only, yielding a total of ~2200 druggable genes.

Conclusions

What do these numbers tell us and how stable can we expect them to be? As discussed already in this article, this approach to annotating the druggable genome estimates the potential maximum size of the playing field for current small-molecule drug design. It does not consider biologicals (where the rules are harder to define), RNAi (not based on protein structure) or future breakthroughs in medicinal chemistry or biology (no crystal ball available).

We have to remind ourselves that these simple sequence comparisons do not allow any immediate far-reaching conclusions about the biological function of a protein. However, sequence based protein-structure prediction certainly pinpoints areas of research that will be highly enriched for 'real' drug targets. Investigating the function of orphan receptors has been very successful in identifying unexpected novel players in important physiological pathways, for example, see Refs [16,17], and the druggable genome will guide

feature

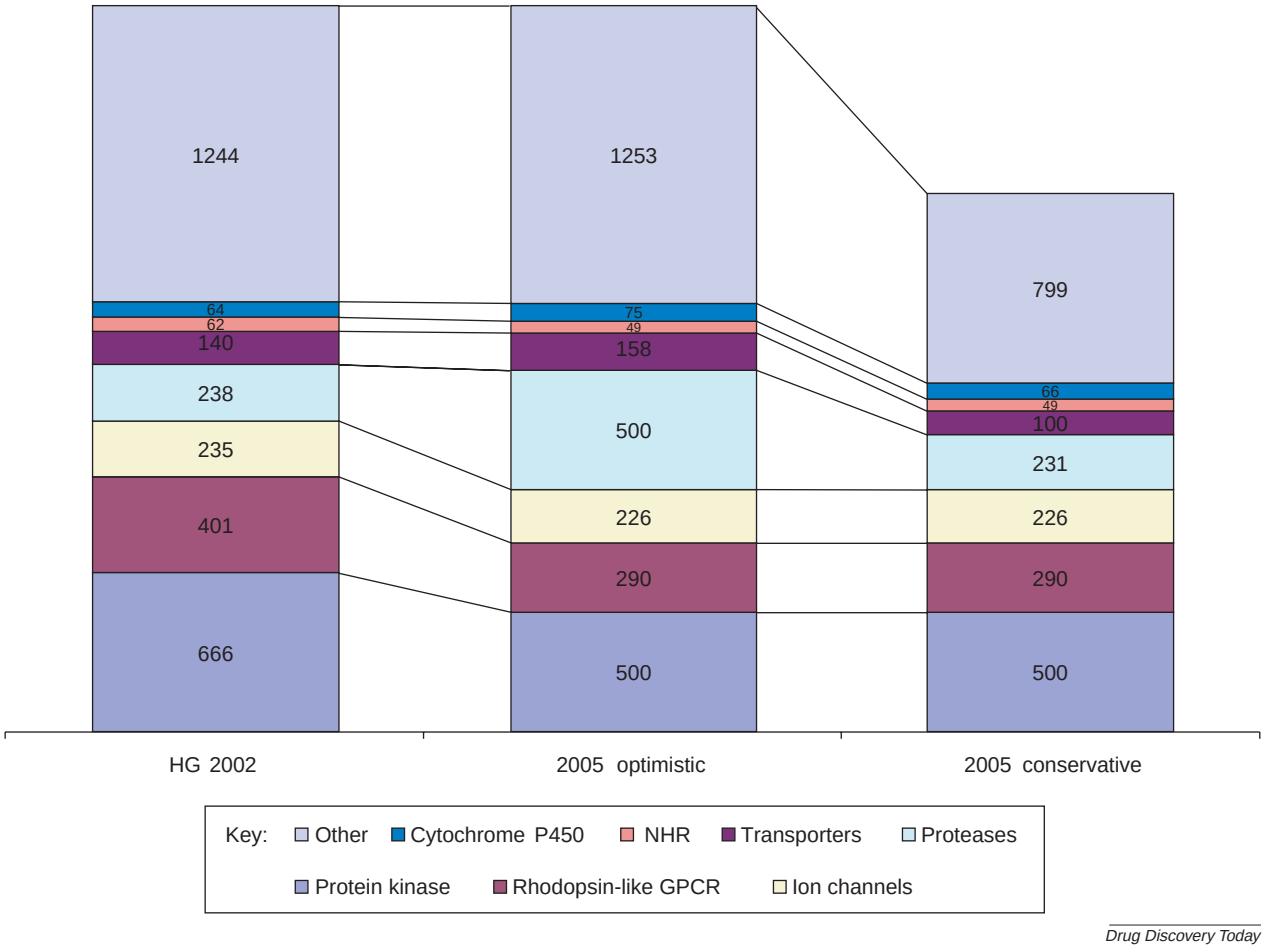

FIGURE 2

Optimistic and conservative estimates of the druggable genome compared with previous predictions [2]. Abbreviations: HG, Hopkins and Groom [2]; NHR, nuclear hormone receptors.

experimental efforts further towards understanding the biology of other potential drug targets.

References

1 International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945

2 Hopkins, A.L. and Groom, C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.* 1, 727–730

3 Hopkins, A.L. and Groom, C.R. (2003) Target analysis: a priori assessment of druggability. *Ernst Schering Res. Found. Workshop* 2003, 11–17

4 Orth, A.P. *et al.* (2004) The promise of genomics to identify novel therapeutic targets. *Expert Opin. Ther. Targets* 8, 587–596

5 Muller, G. (2003) Medicinal chemistry of target family-directed masterkeys. *Drug Discov. Today* 8, 681–691

6 Lander, E.S. (2004) Eric S. Lander. *Nat. Rev. Drug Discov.* 3, 730

7 Fishman, M.C. and Porter, J.A. (2005) Pharmaceuticals: a new grammar for drug discovery. *Nature* 437, 491–493

8 Drews, J. and Ryser, S. (1997) The role of innovation in drug development. *Nat. Biotechnol.* 15, 1318–1319

9 Drews, J. (2000) Drug discovery: a historical perspective. *Science* 287, 1960–1964

10 Hubbard, T. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.* 33, D447–D453

11 Mulder, N.J. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.* 33, D201–D205

12 Bateman, A. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.* 32, D138–D141

13 Manning, G. *et al.* (2002) The protein kinase complement of the human genome. *Science* 298, 1912–1934

14 Fredriksson, R. *et al.* (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* 63, 1256–1272

15 Puente, X.S. *et al.* (2003) Human and mouse proteases: a comparative genomic approach. *Nat. Rev. Genet.* 4, 544–558

16 Messager, S. *et al.* (2005) Kisspeptin directly stimulates gonadotropin-releasing hormone release via G protein-coupled receptor 54. *Proc. Natl. Acad. Sci. U. S. A.* 102, 1761–1766

17 Seminara, S.B. *et al.* (2003) The GPR54 gene as a regulator of puberty. *N. Engl. J. Med.* 349, 1614–1627

18 Hsu, F. *et al.* (2005) The UCSC Proteome Browser. *Nucleic Acids Res.* 33, D454–D458

19 Roth, B.L. *et al.* (2004) Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.* 3, 353–359

20 Morphy, R. *et al.* (2004) From magic bullets to designed multiple ligands. *Drug Discov. Today* 9, 641–651

Andreas P. Russ* and Stefan Lampel
Genetics Unit,
Department of Biochemistry,
University of Oxford,
South Parks Road,
Oxford OX1 3QU,
UK
**e-mail: andreas.russ@bioch.ox.ac.uk*