Properties and identification of human protein drug targets

Tala M. Bakheet and Andrew J. Doig

Received on July 22, 2008; revised on December 19, 2008; accepted on December 30, 2008

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Abstract

Motivation: We analysed 148 human drug target proteins and 3573 non-drug targets to identify differences in their properties and to predict new potential drug targets.

Results: Drug targets are rare in organelles; they are more likely to be enzymes, particularly oxidoreductases, transferases or lyases and not ligases; they are involved in binding, signalling and communication; they are secreted; and have long lifetimes, shown by lack of PEST signals and the presence of N-glycosylation. This can be summarized into eight key properties that are desirable in a human drug target, namely: high hydrophobicity, high length, SignalP motif present, no PEST motif, more than two N-glycosylated amino acids, not more than one O-glycosylated Ser, low pl and membrane location. The sequence features were used as inputs to a support vector machine (SVM), allowing the assignment of any sequence to the drug target or non-target classes with an accuracy in the training set of 96%. We identified 668 proteins (23%) in the non-target set that have target-like properties. We suggest that drug discovery programmes would be more likely to succeed if new targets are chosen from this set or their homologues.

Contact: andrew.doig@manchester.ac.uk

Supplementary Information: Supplementary data are available at Bioinformatics online.

1 Introduction

Drug discovery typically begins by the identification and validation of a potential target which is then subject to high-throughput screening against a library of drug-like compounds or to rational drug design. It is vital to have as much evidence as possible to support a target choice before investing more resources in the target. Good drug targets share several features: involvement in a crucial biological pathway; distinction from any previously known target; functionally and structurally characterized; and druggable (capable of binding to small molecules, implying the presence of a binding site). When searching for novel drug targets, candidates can be assessed according to how many of these features they have, as well as participation in a biological process critical to disease.

Structure-based analysis has led to the concept of ‘druggability’ which is used to describe proteins that possess protein folds that favour interactions with drug-like chemical compounds (Hopkins and Groom, 2002; Keller et al., 2006; Orth et al., 2004; Russ and Lampel, 2005). Many proteins are druggable according to their structure, but their binding will not lead to the therapeutic benefit. Druggable proteins are thus not necessarily drug targets. Drug binding sites are expected to have certain structural and physicochemical properties with a high affinity for site-specific binding (Zheng et al., 2006). Analysis of binding site properties can be used to predict protein druggability (Cheng et al., 2007; Nayal and Honig, 2006). Such approaches require a protein structure, however, which is not available for most proteins, particular those found in membranes.

At present, the most frequent protein targets for which successful drugs have been developed include proteases, kinases, G protein-coupled receptors (GPCRs) and nuclear hormone receptors (Drews, 2000; Imming et al., 2006). GPCRs (23%) and enzymes (50%) represent the most important target classes of proteins for drug discovery (Zheng et al., 2006). Hopkins and Groom (Hopkins and Groom, 2002) identified 120 successful proteins as rule-of-five compliant, oral drug targets for approved drugs, using sequence motifs, where the two largest molecular target classes were GPCRs and enzymes. They estimated that around 3000 human proteins were druggable. Russ and Lampel (Russ and Lampel, 2005) concluded that there are 2000–3000 druggable proteins in humans, with fewer rhodopsin-like GPCRs and protein kinases, but more proteases than in previous work. Overington et al. (2006) surveyed earlier reports to propose a consensus number of 324 drug targets for all classes of approved therapeutic drugs. Lauss et al. (2007) reported properties of 392 human drug targets, assessing their biochemical functions, involvement in biological processes, cellular locations, biological pathway associations, tissue distribution and chromosome distribution. Zheng et al. (2006) reported additional properties of new druggable proteins, namely: membership of a target family, sufficient sequence variation to allow differential binding of small molecules, a small number of similar proteins outside its family, involvement in no more than two pathways and presence in no more than two tissues. While it is possible to search genomes for novel members of known families, a more adventurous strategy is to seek drug targets from entirely new classes.

One way of identifying new potential drug targets is to analyse the human genome on the basis of sequence homology to the known therapeutic drug targets or possession of the same sequence motifs. More information is available from protein structural data. While structural information can be used to identify homologues that are not apparent from sequence data, structures can also be used to assess their potential ability to bind drug-like molecules.
(Hajduk et al., 2005). Drug targets may also be identified via side effect similarity (Camillos et al., 2008).

Here, we report a number of properties of human drug targets that differ significantly from non-drug targets and apply these to identify new potential drug targets. Previous work on categorizing drug targets has considered biological classes identified by sequence similarity. Here, we use a much wider range of sequence properties to help characterize drug targets. Understanding what makes a good drug target will aid with the process of target selection. As Lipinski and colleagues (2001) found rules to help identify good drugs, the research described in this article gives trends to help identify good drug targets.

2 METHODS

2.1 Dataset

The drug target dataset was downloaded from DrugBank (Wishart et al., 2008; http://redpoll.pharmacy.ualberta.ca/drugbank/download.htm). The approved drug target set (Version 1.0) contains around 500 proteins. The frequency of each amino acid in each protein is divided by protein length to give percentage frequencies. Amino acids were grouped as follows: tiny (A, C, G, S and T); small (A, B, C, D, G, N, P, S, T and V); aliphatic (I, L and V); aromatic (F, H, W and Y); polar (D, E, H, K, M, P, V, W and Y); non-polar (A, C, F, G, I, L, M, P, V, W and Y); polar (D, E, H, K, N, Q, R, S, T and Z); charged (B, D, E, H, K, R and Z); basic (H, K and R). Hydrophobicity for drug targets and non-target proteins was calculated as the sum of hydrophobicity values using the Kyte and Doolittle index (Kyte and Doolittle, 1982), divided by the number of residues in each of the protein sequences. The Pepstats program (http://emboss.cbr.nrc.ca/cgi-bin/emboss/pepstats) was used to output all protein sequences. The Pepstats program (http://www.cbs.dtu.dk/services/NetPhos/) was used for identification of phosphorylation sites with a sensitivity >70% (Blom et al., 1999).

The NetOglyc program (http://www.cbs.dtu.dk/services/NetO Glyc/) was used for identification of O-glycosylation sites using neural networks (Julenius et al., 2005). The NetNglyc program (http://www.cbs.dtu.dk/services/NetNGlyc/) was used for identification of N-glycosylation sites (Jensen et al., 2003).

2.2 Simple sequence properties

The frequency of each amino acid in each protein is divided by protein length to give percentage frequencies. Amino acids were grouped as follows: tiny (A, C, G, S and T); small (A, B, C, D, G, N, P, S, T and V); aliphatic (I, L and V); aromatic (F, H, W and Y); non-polar (A, C, F, G, I, L, M, P, V, W and Y); polar (D, E, H, K, N, Q, R, S, T and Z); charged (B, D, E, H, K, R and Z); basic (H, K and R). Hydrophobicity for drug targets and non-target proteins was calculated as the sum of hydrophobicity values using the Kyte and Doolittle index (Kyte and Doolittle, 1982), divided by the number of residues in each of the protein sequences. The Pepstats program (http://emboss.cbr.nrc.ca/cgi-bin/emboss/pepstats) was used to output all protein sequence information statistics including isolectric points (pI values), and number of positively charged and negatively charged amino acids. A PERL script was written to extract protein sequence features and calculate the mean frequencies.

2.3 EC number, glycosylation and phosphorylation

Primary Enzyme Commission (EC) numbers, N- and O-glycosylation and phosphorylation of Ser, Thr and Tyr were taken from SWISS-PROT annotations. Primary EC numbers were taken from the DE line that contains the EC classification, where primary EC numbers 1–6 are oxidoreductase, transferase, hydrolase, lyase, isomerase and ligase, respectively. N- and O-glycosylation was taken from the reference position line as RP GLYCOSYLATION, while phosphorylation of Ser, Thr and Tyr was extracted from the RP PHOSPHORYLATION line.

Two programs were used to predict phosphorylation sites: Scansite (Obenauer et al., 2003) is a computational tool built on experimental binding and/or substrate information that predicts protein–protein interactions and identifies phosphorylated sites using a position-specific scoring matrix. Sequence files were uploaded onto http://stjuderesearch.org/scansite/ with a stringent level set to high to increase the reliability of the true positives. The NetPhos method (http://www.cbs.dtu.dk/services/NetPhos/) uses neural networks based on sequence and structural information to predict phosphorylation sites with a sensitivity >70% (Blom et al., 1999).

The NetOglyc program (http://www.cbs.dtu.dk/services/NetO Glyc/) was used for identification of O-glycosylation sites using neural networks (Julenius et al., 2005). The NetNglyc program (http://www.cbs.dtu.dk/services/NetNGlyc/) was used for identification of N-glycosylation sites (Jensen et al., 2003).

2.4 Gene Ontology terms

The Gene Ontology (GO) Consortium provides a controlled vocabulary to describe gene function (Ashburner et al., 2000), using the following three organizing principles: (i) biological process; (ii) molecular function; and (iii) cellular component. The GO database was downloaded from http://www.geneontology.org/GO.downloads.shtml, dated September 2006. Another database was downloaded from http://www.geneontology.org/doc/GO.terms_and_ids to map all GO term IDs to their names and ontologies. GO ID annotations are in SWISS-PROT in the DR line. A PERL program was coded to extract the GO IDs for each dataset and to parse a GO file and write child–parent relationships as a full path for every GO term and to output the highest level and next highest level for every path as levels 1 and 2. The numbers of molecular function entries was 207, biological process entries was 299 and cellular component entries was 135. If a GO term ID was present more than once at level 1 or 2, it was only counted once at that level.

2.5 Subcellular location

Subcellular location is annotated in SWISS-PROT as SUBCELLULAR LOCATION, and is mainly found in the comment lines (CC). Some annotations contain the terms probable/potential/similarity; the analysis was performed using all these terms. The WoLF PSORT program (http://wolfsort.seq.cbrc.jp/) (Horton et al., 2007) was used to predict protein subcellular localization. It makes predictions based on known sorting signal motifs and sequence features such as amino acid content. The output report of WoLF PSORT was analysed by extracting the location of the highest predicted location for both datasets. The WoLF PSORT program was used as it can make a prediction on any sequence.

2.6 Signal peptide cleavage

Signal peptide is annotated in SWISS-PROT as FT SIGNAL in the feature table (FT) line. Some annotations contain the terms probable/potential/similarity and hence are expected to be less reliable. We therefore performed the analyses with and without entries with these annotations. The SignalP program (http://www.cbs.dtu.dk/services/SignalP/) was used to perform signal peptide prediction (Bendtsen et al., 2004).
2.7 Transmembrane helices
Transmembrane helices are annotated in SWISS-PROT as FT
TRANSMEM in the feature table line (FT). The TMHMM method
(http://www.cbs.dtu.dk/~mmk/tmhmm/) (Krogh et al., 2001) was used
to predict the location and orientation of α-helices in membrane-spanning
proteins.

2.8 PEST regions
PEST sequences are hydrophilic regions of 12 or more amino acids that
contain at least one P, one E or D, and one S or T. The PESTfind program
(http://emboss.cbr.nrc.ca/cgi-bin/emboss/pestdetect) was used to output all
poor and potential PEST protein sequences. The criterion to be a potential
PEST sequence of real biological interest is to have a ‘Valid’ PEST motif.
Valid PEST motifs have a value above the threshold score (5.0), while motifs
below the threshold score are considered ‘poor’.

2.9 Low-complexity regions
The SEG program (Wootton and Federhen, 1993) was downloaded from
ftp://ftp.ncbi.nih.gov/pub/seg/). It is used to mask composition-biased
regions in the query, based on a statistical approach.

2.10 Secondary structure
The JPred program (http://www.compbio.dundee.ac.uk/~www-jpred/) (Cuff
et al., 1998) was used to predict the percentages of α-helix and β-sheet in 148
drug targets and a sample of 1064 non-targets. Accurate secondary structure
could have been found from crystal structures. However, most proteins do
not have structures, particularly if they are membrane proteins, which are over-
represented in the drug target set. We therefore used a secondary structure
prediction program, rather than using structures.

2.11 Machine learning
Support vector machines (SVMs) (Vapnik, 1995), using a radial basis
function and the WEKA package (Witten and Frank, 2005) were used
to make a classifier that can distinguish targets from non-targets, using
features that can be calculated from any protein sequence. The features we used were:
amino acid compositions, length, hydrophobicity, SignalP,
PEST, NetOglyc Ser, NetOglyc Thr, low-complexity regions, α-helix, β-
sheet, transmembrane helices and pf. A scaling scheme was used for every
vector by restricting all entries to be between 0 and 1, by calculating for
every feature (X−Min)/(Max−Min) where X is the feature score, and Min
and Max are the minimum and maximum values of X in the set.

The SVM was trained using 5-fold cross-validation on a dataset of 99
targets and 103 non-targets. Similarly sized datasets were used to avoid a
bias to the larger set. Two parameters were varied to maximize the accuracy:
the error penalty (C) for an incorrect prediction and the radial basis function
parameter (γ) which controls how smooth the boundary is between the target
and non-target areas. The optimal parameters for C and γ were 0.6 and
0.92, respectively. This model had an accuracy of 89.4% using 32 features.
A genetic algorithm was used to optimize feature use. The initial population
was randomly generated. For each chromosome, the fitness function is represented by the accuracy of the SVM, as the number
of protein correctly classified. The calculated accuracy scores were used to
select 20% of chromosomes to survive to the next generation. These survivors
are copied with mutations to produce the next generation, with a 10%
probability of mutating each 0 to 1 or vice versa. The process was terminated
after 15 generations. Removing two features in this way (Trp content and
SignalP) improved the accuracy to 96.5%. The model was verified on a
validation set of 49 targets and 45 non-targets that was not used in
optimizing the SVM. The validation set classified 5 targets and 5 non-targets
incorrectly, with 84 correctly assigned, giving an accuracy of 84/94 = 89%.

3 RESULTS
3.1 Simple sequence properties
Figure 1 shows the amino acid frequencies for the two datasets.
There is a trend for drug targets to have more non-polar amino acids
than non-drug targets. The differences for Val and Gly were significant, with P-values of 0.001 and 2.6 × 10−7 respectively.
Targets had significantly higher proportions of aromatic (11.3% versus 10.6%; P = 0.00016) and non-polar (56.3% versus 53.6%;
P = 7 × 10−8) amino acids, and lower proportions of positive (12.8% versus 14.2%; P = 1.6 × 10−5), polar (43.7% versus 46.4%;
P = 1.7 × 10−7), charged (24.0% versus 25.6%; P = 0.001) and
basic (12.8% versus 14.2%; P = 2 × 10−5) amino acids. Differences in proportions of negative amino acids were not significant. Target
proteins tend to be longer than non-targets, with mean numbers of
577 and 523 residues (P-value = 9.67 × 10−5).

The mean hydrophobicity for the amino acids in the drug targets
was higher than in the non-drug targets, where −97 and −190.7
were the calculated hydrophobicity values for the drug targets and
non-drug targets, respectively, with a significant P-value of 0.002.
These results confirmed the amino acid composition results, where
drug target proteins are more hydrophobic and less polar.

The mean drug target pI was lower than the mean non-target protein pI with values of 6.95 and 7.44 for the target and non-targets,
respectively, with a P-value of 0.005. This was affected by the higher
mean frequency of positively charged amino acids in non-targets
over targets.

3.2 Enzyme class
Figure 2 shows the distributions of primary Enzyme Class numbers
in drug targets and non-drug targets. Oxidoreductases and lyases are
more frequent as targets compared to non-drug targets, while ligases,
isomerases and hydrolases are disfavoured as targets. Ligases, in
particular, appear to be strongly disfavoured as targets, being three
times more frequent as non-targets, though the numbers are small.
While hydrolases are frequent drug targets, this is still lower than
their occurrence in non-drug targets.
terms favoured for non-drug targets include physiological process, molecular metabolism and regulation of physiological process, though there are strong disfavour organelles. For biological processes, terms favoured for drug targets to be extracellular or membrane bound and to disfavour organelle, organelle part, intracellular part, intracellular, intracellular part and cell fraction. Terms favoured for non-drug targets include extracellular region part, extracellular region, membrane, membrane part and cell fraction. Terms disfavour non-drug targets include transcription regulator activity and nucleic acid binding. This indicates that drug targets are strongly disfavoured as drug target locations, with non-drug targets frequently located in the nucleus, mitochondria and Golgi apparatus, while drug targets are rare or unknown in these locations.

3.3 Sub-cellular location

Figure 3 shows the distributions of sub-cellular locations for drug targets and non-drug targets. Strong trends are seen, with over half of drug targets found in membranes. Drug targets are also more likely to be extracellular than non-drug targets. Organelles are strongly disfavoured as drug target locations, with non-drug targets frequently located in the nucleus, mitochondria and Golgi apparatus, while drug targets are rare or unknown in these locations.

3.4 GO terms

The distributions of GO terms at levels 1 and 2 for molecular function, cellular component and biological process are shown in Supplementary Figure 1. For molecular function, terms favoured for drug targets include catalytic activity, signal transducer activity, transporter activity, receptor activity, lyase activity, ion transporter activity and channel or pore class transporter activity. Terms disfavour non-drug targets include transcription regulator activity and nucleic acid binding. This indicates that drug targets are frequently enzymes and involved with binding and signalling. For cellular component, terms favoured for drug targets include extracellular region part, extracellular region, membrane, membrane part and cell fraction. Terms disfavour non-drug targets include organelle, organelle part, intracellular part, intracellular, intracellular organelle and membrane-bound organelle. This shows preferences for drug targets to be extracellular or membrane bound and to disfavour organelles. For biological processes, terms disfavour for drug targets include cellular process and cell communication. Terms disfavour for non-drug targets include physiological process, metabolism and regulation of physiological process, though there are no very strong trends. We submitted the target dataset to GOstat

(Reissbarth and Speed, 2004), which outputs significant GO terms present within the set (Supplementary Table 2). Lauss et al. (2007) reported similar results.

3.5 Signal peptide cleavage

Signal peptide sequences control protein entry to the secretory pathway for export and directs the protein to organelles such as nucleus, mitochondrial matrix, endoplasmic reticulum, chloroplast and peroxisome. Signal peptide motifs are more frequent in drug targets (50/148 = 33%) than non-drug targets (563/3573 = 15%) with a P-value of 0.06. When entries containing the terms probable, potential or similarity are excluded, the difference was even more apparent, as the frequency in drugs targets was 33/148 = 22%, while the non-drug target frequency was 184/3573 = 5%, with a P-value of 0.007. Results from the Signal P prediction program supported these observations, confirming that signal peptide motifs are more frequent in drug targets (65/148 = 43%) than non-drug targets (909/3573 = 25%), with a P-value of 4.01 × 10^{-5}. Drug targets are thus more likely to be secreted.

3.6 Transmembrane helices

Annotations showed that 43% (64/148) of targets contain transmembrane helices, while 20% (715/3573) of non-drug targets contain transmembrane helices, with a significant P-value of 0.02. This confirms the sub-cellular location and GO data indicating that drug targets are often found in membranes. The TMHMM algorithm confirmed that target proteins had more transmembrane helices than non-targets (323/148 = 2.18 helices per protein in targets; 769/3573 = 0.77 helices per protein in non-targets; P-value = 4.9 × 10^{-7}), though the number of helices was much higher using the prediction program than annotations.

3.7 Phosphorylation

The reversible phosphorylation of proteins plays an essential role in many fundamental cellular functions, including survival, differentiation, structural organization, cell division, cell differentiation and stress responses. SWISS-PROT annotations had 8% (13/148) of target proteins and 10% (386/3573) of non-target proteins phosphorylated. As approximately one-third of proteins is believed to be phosphorylated, these annotations are likely to be missing most cases. Predictions of phosphorylation using the ScanSite and NetPhos programs were therefore also made.

Table 1 shows the results from using the ScanSite and NetPhos programs on the two datasets. The numbers submitted to each program differ, as some sequences would not run and the servers had different limitations. The results from the two programs differ, with NetPhos predicting far more phosphorylation sites than ScanSite. Nevertheless, both programs predict that drug targets are less likely to be phosphorylated than non-drug targets, though these differences are small. This may be because drug targets are more likely to be membrane-bound, where much of their surface is inaccessible to kinases.

3.8 Glycosylation

Table 2 shows the predicted numbers of O- and N-glycosylation sites. O-glycosylation is more frequent in targets, while N-glycosylation is more frequent in non-targets. Many O-glycosylated
proteins form part of the extracellular matrix or mucosal secretions, neither of which is expected to contain many drug targets. N-glycosylation is linked to resistance to degradation, implying that drug target proteins have a longer lifetime in vivo, as well as protein folding and assembly (Peter-Katalinić, 2006).

3.9 PEST regions
PEST sequences are implicated in the regulation of protein degradation. Potential PEST motifs in drug target proteins were less frequent than in non-target proteins, with values of 52/148 = 0.35 and 2516/3573 = 0.70 sites per protein for drug targets and non-targets, respectively (P-value = 0.001). This result suggests that drug target proteins have a longer lifetime when compared to non-target proteins, in agreement with the N-glycosylation results.

3.10 Low-complexity regions
Low-complexity segments in protein sequences are commonly short repeats and interspersed regions. Low-complexity regions in drug target were slightly less frequent in the drug target proteins than in non-target proteins with values of the mean frequencies for drug targets of 2.35 (349/148) and non-targets 2.43 (8715/3573). This small difference was not significant (P-value = 0.88).

3.11 Secondary structure
The predicted α-helix contents in the target and non-target sets were very similar (target: 32.0% from 148 sequences; non-target 30.8% from 1064 sequences; P-value = 0.08). β-Sheets were predicted to be more frequent in targets than non-targets (target: 16.8% from 148 sequences; non-target 14.2% from 1064 sequences; P-value = 1.2 × 10^-5).

3.12 Membrane and non-membrane proteins
A major difference between the target and non-target proteins is that target proteins are more likely to be found in membranes. It is therefore possible that the differences we find simply arise from properties varying with the likelihood of a protein being in a membrane, rather than being a drug target. We therefore divided each set into membrane and non-membrane proteins. We compared membrane targets with membrane non-targets, and non-membrane targets with non-membrane non-targets. Many properties were still significantly different between targets and non-targets, including signal P sequences, pI, length, hydrophobicity, transmembrane helices, N-glycosylation, phosphorylation, low-complexity regions and PEST sequences (Supplementary Table 3). In addition, SVMs trained to distinguish targets from non-targets with membrane proteins or within non-membrane proteins were also accurate. We are therefore confident that these properties reflect real differences between targets and non-targets, rather than simply differing likelihoods of being a membrane protein.

3.13 Human drug target rules
Our results can be summarized and simplified into a set of eight properties that are desirable for a human drug target protein (Table 3). A drug target should have as many of these properties

<table>
<thead>
<tr>
<th>Property</th>
<th>P (Target)^a</th>
<th>P (Non-target)^b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydrophobicity &gt; −142.4</td>
<td>0.66</td>
<td>0.60</td>
</tr>
<tr>
<td>Length &gt; 550 amino acids</td>
<td>0.39</td>
<td>0.29</td>
</tr>
<tr>
<td>SignalP motif present</td>
<td>0.45</td>
<td>0.27</td>
</tr>
<tr>
<td>No PEST motif</td>
<td>0.21</td>
<td>0.35</td>
</tr>
<tr>
<td>More than two N-glycosylated amino acids</td>
<td>0.52</td>
<td>0.38</td>
</tr>
<tr>
<td>Not more than one O-glycosylated Ser</td>
<td>0.16</td>
<td>0.24</td>
</tr>
<tr>
<td>Mean pI &lt; 7.2</td>
<td>0.37</td>
<td>0.51</td>
</tr>
<tr>
<td>Membrane location</td>
<td>0.49</td>
<td>0.24</td>
</tr>
</tbody>
</table>

^aProbability that a drug target protein has this property. ^bProbability that a non-drug target protein has this property.
We used the sequence properties and support vector machines to build a classifier that can distinguish between the target and non-target sets. The most accurate model, after feature selection using a genetic algorithm, used the following sequence features: all amino acid compositions except Trp, length, hydrophobicity, PEST, NetOglyc Ser, NetOglyc Thr, low-complexity regions, α-helix, β-sheet, transmembrane helices and pl. This had an accuracy of 89%. While the trends for desirable properties listed in Table 3 are often weak, when combined they give a predictor of high accuracy. Features were chosen to be determined from any protein sequence, leaving out attributes that are not available for some sequences, such as those based on experimental annotation or structure. The model was applied to the entire non-target dataset. As discussed above, this set contains proteins that have not yet been found to be drug targets, and therefore contains many which may be targets of the future. Our goal was to identify the proteins within this set that resemble the known targets. Ideally the entire human proteome would be run, but this would have taken too much time. Some of the servers used to calculate sequence features are limited to 50 sequences per submission, for example. It was not possible to run all the sequences, such as those over 800 amino acids, so only 2923 out of the 3573 non-targets were assigned. This does give nearly complete coverage of the human proteome, however, as even if a particular sequence was not analysed, one of its homologues probably was.

We found that 668/2923 (23%) of the non-targets were misclassified within the target set (Supplementary Table 4). In other words, 23% of human proteins have target-like properties. Some of these have been annotated by primary EC number, giving 17 oxidoreductases, 12 transferases, 44 hydrolases, 6 lyases, 5 isomerases and no ligases. These are similar proportions to the known targets (Fig. 2), though with rather more hydrolases. The lack of ligases assigned to be targets is consistent with known target data, in particular. The distance from the SVM hyperplane gives a measure of the accuracy of the assignments. A positive hyperplane distance is an assignment to a drug target and a negative hyperplane distance is an assignment to a non-drug target. Further the score is from zero, the more reliable is the prediction.

4 DISCUSSION

Numerous sequence properties show significant differences between drug target proteins and non-drug targets. Drug targets tend to be found in membranes or are extracellular; they are rare in organelles; they are more likely to be enzymes, particularly oxidoreductases, transferases or lyases and not ligases; they are involved in binding, signalling and communication; they are secreted; and have long lifetimes, shown by lack of PEST signals and the presence of N-glycosylation. Using imperfect prediction programs, rather than annotations, adds strength to these conclusions, since the P-values from imperfect predictors will be an upper bound on the true P-value (Supplementary Material Document 1). We suggest that these properties should be considered when selecting a new target for drug development, in addition to the protein’s participation in a pathway of biological interest. We expect that taking these considerations into account will reduce the likelihood of a drug discovery programme failing, even if highly active, rule-of-five compliant drug with excellent pharmacokinetic properties are being used. For example, a protein that is found in the nucleus, contains PEST motifs, lacks N-glycosylation, is short and acts as a ligase would be expected to be a poor target compared to an extracellular oxidoreductase with a signal peptide motif, N-glycosylation and no O-glycosylation or phosphorylation. Of course, caution should be exercised in using these rules if future drug discovery moves into new areas, such as perturbing protein–protein interactions or inhibiting protein aggregation.

The features were used as inputs for a support vector machine, trained to distinguish between targets and non-targerts. A high accuracy was achieved after optimization of SVM parameters and feature selection. The algorithm was run on 2923 non-redundant sequences within the non-target set and 668 were assigned as targets. These have sequence properties that are more similar to those within the target training set, than within the non-target training set. We suggest that future drug discover programmes will stand a higher chance of success if targets are chosen from these 668 proteins or one of their homologues.

ACKNOWLEDGEMENTS

We thank Roland Ennos for advice on statistics and Paul Dobson for helpful discussions.

Funding: Biomolecular Research Unit at King Faisal Specialist Hospital & Research Centre (KFSH&RC) for a PhD Studentship (to T.B.).

Conflict of Interest: none declared.

REFERENCES

Human protein drug targets


