J Comput Aided Mol Des (2009) 23:583–592
DOI 10.1007/s10822-009-9279-y

Structural models in the assessment of protein druggability based on HTS data

Anvita Gupta · Arun Kumar Gupta · Kothandaraman Seshadri

Received: 27 January 2009 / Accepted: 12 May 2009 / Published online: 29 May 2009

Abstract Insights on the potential of target proteins to bind small molecules with high affinity can be derived from the knowledge of their three-dimensional structural details especially of their binding pockets. The present study uses high-throughput screening (HTS) results on various targets, to obtain mathematical predictive models in which a minimal set of structural parameters significantly contributing to the hit rates or the affinity of the protein binding pockets for small molecular entities, is identified. An emphasis is given to focus on target variation aspect of the data by consideration of commonly tested compounds against the HTS targets. We identify 'four-parameter' models with R^2, R^2_{adj}, *SEE*, and *LOO* q^2 values of 0.70, 0.60, 0.27 and 0.50, respectively, or better. We demonstrate through cross-validation exercises that our regression models apply well on varied data sets. Thus we can use these models to estimate hit rates for HTS campaigns and thereby assign priority to drug targets before they undergo such resource intense experimental screening and follow-up.

Keywords Druggability · Target identification · Target validation · Protein binding pockets · High throughput screening · Hit rates

Abbreviations

CCT	Common compounds tested
FDA	Food and drug administration (U.S. Department of Health and Human Services)
HR	Hit rate
HTS	High-throughput screening
IQR	Inter quartile range
LOO	Leave-one-out
MWSS	Model without site score
NME	New/novel molecular entity
NMR	Nuclear magnetic resonance
SEE	Standard error of estimate

A. Gupta · A. K. Gupta · K. Seshadri (✉)
AstraZeneca India Private Limited, Avishkar Building, Kirloskar Business Park, Bellary Road, Hebbal, Bangalore 560024, India
e-mail: kothandaraman.seshadri@astrazeneca.com

Introduction

Key attributes to an ideal druggable target would include it being a causative factor in a disease and its ability to be inhibited by small molecules that are orally bio-available [1, 2]. Despite genomic initiatives and our understanding of various pathways, target innovation has been very slow. The paucity in finding such targets for various therapies has been a subject of intense deliberation considering huge jumps in Pharma R&D expenditure versus the resultant number of novel drug targets and NMEs. In the period of 1993–2003, of more than 300 NMEs approved by FDA, only less than 6% were known to modulate novel protein activities [3, 4] and on an average, only three new drug targets are addressed per year with synthetic drugs [1]. Although this can be mainly attributed to attritions in the late stages of drug discovery, careful considerations should be made at early stages of target identification and validation. Wrong choice of targets ultimately becomes a burden on resources even at initial screening stages. Drug action is considerably more complex than the display of binding affinity to a drug target, nevertheless a 'target filter' based on its propensity to be druggable would be a desirable tool in prioritizing, especially when dealing with equally promising but marginally distinguishable targets.

Computational approaches to describe a target's druggability in general range from a broader genome-wide level analysis to structural characterization of binding sites and further to specific parametric mathematical regression models [2, 5–8].

Protein molecules as a class have been attractive drug targets owing to the ability to bind specific inhibitors [1, 9]. The physicochemical properties and pharmacophores of these inhibitors thereby complement the binding sites of biological molecules. Akin to Lipinski *rule-of-five* [10, 11], which describes physicochemical descriptors of small molecules to predict their oral availability or 'drug-likeness', one could think of structural descriptors that define the binding site for these drugs-like molecules. Thus, drug-binding sites should also have certain structural and physiochemical properties to accommodate high affinity site-specific binding and subsequent regulation of protein activity by drug like molecules [12].

To computationally explore the structural basis of druggability of a protein, it is important to consider why the binding pocket within a target molecule assumes significance. Most often the pockets are predefined and exist out of functional necessity and to impart unique character to the protein molecules. Otherwise an upkeep of an accidentally formed pocket would invite thermodynamic penalty due to factors such as solvent exposure of hydrophobic groups or to avoid inappropriate modulation by cellular metabolites [9]. The knowledge that the structure of a bio-macromolecule is decisive in its function stems not only from its residue composition, but also from the residues' spatial juxtaposition especially in the regions forming suitable loci of functional groups that could attract, bind and catalyze small molecules.

The analysis of physicochemical properties of a putative drug-binding pocket of a target protein can provide insights into the target's druggability. In a recent study Hajduk and coworkers [5] have pioneered the computational assessment of molecular druggability by comparing hit rates from NMR fragment screening with properties derived from the 3D structure of the corresponding targets. This empirical approach has set the tone for computationally assessing molecular druggability. In their work they could derive a quantitative relationship between protein binding site descriptors such as pocket shape or hydrophobicity and the NMR hit rates.

HTS as a hit identification technique is well established in pharmaceutical industry [13–15] and has resulted in a vast body of screening data. It is worthwhile analyzing this data in a manner complementary to the NMR screening studies done by Hajduk and coworkers, in order to derive predictive models that adequately describe the ability of a protein-binding pocket to bind small molecules. In this context, HTS data can be very useful. In this study we used in-house HTS data for 22 target proteins where structural information is available to build a quantitative, regression relationships to obtain relative weights of the binding site descriptors that describe binding site affinity. Given that the HTS and NMR screening campaigns are integral part of discovery efforts [16] and that not many targets can be subjected to this resource intense scheme, our aim is to derive indices that can have implications in the prioritization of drug targets for such campaigns.

Materials and methods

Datasets

The main drivers in our choice for datasets are (1) Appropriate experiment that represents the affinity of target's binding pockets for a range of small molecules, (2) access to the experimental data, (3) diversity profile of the collection of small molecules screened and (4) access to 3D structural information of the target preferably in complex with a ligand. Based on these drivers, we collated in-house HTS data of various eukaryotic and prokaryotic drug targets. This led to a compilation of 34 in-house targets for the study. The hit rates associated with the target proteins would be influenced by the nature and the size of the compound library. The targets in our study have been screened with a library of compounds ranging from about 50,000 to a million (a significant 20-fold variation). Thus to overcome this limitation, instead of calculating the hit rates based on the entire collection of compounds screened on these targets, we derived a subset of compounds that were screened commonly in all target cases (Fig. 1). The set of common compounds *tested* (CCT set) provides us a way of looking at target specific variations. We arrived at a training set of 22 targets for which a reasonably large CCT set (37275 compounds) could be found. In order to rule out the possibility of CCT compounds being promiscuous, a

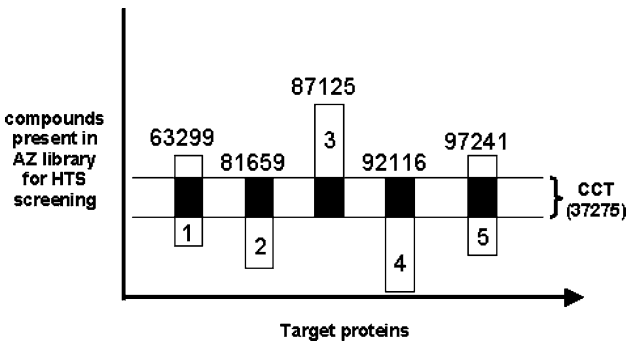

Fig. 1 A conceptual representation of extraction of common compounds tested set (*black*) of in-house HTS data for five different targets

Table 1 Training dataset comprising of 22 target proteins and their HTS hit rates

Common compounds 37275			
S. no	Targets	Hit rate[a] (%)	Log (hit rate)
1	Ligase	0.06	−3.2492
2	Nucleotide kinase-1	0.14	−2.85541
3	Racemase-1	0.24	−2.61718
4	Acyltransferase	0.30	−2.5222
5	Hydrolase, pyrophosphorylase, polymerase	0.34	−2.47451
6	Racemase-2	0.43	−2.3619
7	Phosphooctulonate synthase	0.47	−2.3259
8	Transferase	0.50	−2.29726
9	Topoisomerase	0.65	−2.1876
10	Dual specificity protein kinase	0.76	−2.11657
11	Nucleotide kinase-2	0.79	−2.10013
12	Cysteine endopeptidase-1	0.98	−2.00675
13	Cysteine endopeptidase-2	1.10	−1.95863
14	Kinase-1	1.37	−1.86215
15	Metalloendoproteinase	1.66	−1.77973
16	Hexokinase	1.79	−1.74599
17	Dehydrogenase-1	1.82	−1.73891
18	Phosphodiesterase	1.95	−1.70929
19	Protein tyrosine kinase	2.18	−1.6624
20	Kinase-2	2.23	−1.65234
21	Serine endopeptidase	2.58	−1.58824
22	Dehydrogenase-2	3.85	−1.41426

Hits represent the number of actives among the CCT set and hit rate is the corresponding ratio of hits

[a] The compounds that displayed 'read-out' above 2IQR (Inter Quartile Range) of the median were termed hits. This establishes a confidence level of above 95%

quantity A/T ratio was computed for these compounds. The A/T ratio is the ratio of the number of assay screens in which the compound exhibited activity to the total number of screens in which the compounds were tested. The number of screens that these compounds were subjected to ranged from 251 to 425 and for a majority of compounds of CCT (94.5%), the A/T ratio fell under 0.04. Hence these compounds are expected to be non-promiscuous. The HTS data for these targets had a modest diverse range of hit rates (Table 1).

While maintaining firmer guidelines for a training dataset, validation test data sets can possibly come through a variety of sources. There have been quite a few undertakings to compile protein-ligand interaction datasets that have been used widely, particularly in the field of benchmarking many molecular docking programmes. These include Protein Ligand Database [17] Protein–Protein Interaction Thermodynamic Database [18], Binding DB [19], Ki dataset for protein targets [20] and Böhm's dataset of 82 protein–ligand complexes of known 3D structure and binding constant Ki [21]. While we anticipate that a prediction strategy to be applicable over a disparate sources of data on ligand binding, it is imperative that it is consistent with the source of origin—in our case this being HTS data. After removal of the training set from the 34 targets originally considered, we were left with HTS test dataset comprising of 12 targets, for the later purpose of validating the models. The training set of 22 targets was chosen since the targets contained a reasonable CCT size. As CCT is expected to represent the actual corporate collection, the size of it needs to be significant enough. It is to be noted that the inclusion of any of these 12 proteins into the training set, would have considerably reduced the size of CCT. The resulting test data set is dominated by kinase class—primarily out of the process outlined just above rather than a conscious selection. The hit rates for the targets within this test data set were computed as the number of actives divided by the total number of compounds screened. The logarithm of hit rates ranged between −1.24 and −2.44 (Table 2).

Structure based parameters

The ligand binding pockets (largest ligand, in case of more than one ligands present) in the crystal structures of the targets were considered for the analysis. In few cases where we needed to deal with an uncomplexed structure, pocket with the highest *SiteScore* (a pocket property calculated by *Schrödinger SiteMap* [22]) was considered. *SiteScore* is the weighted sum of *SiteMap* calculated binding pocket

Table 2 List of 12 HTS test dataset proteins along with their experimental hit rates derived from high throughput screening

S. no	Protein	Compounds tested	Hit rate (%)	Log (hit rate)	Pred HR[a] model 3	Pred HR[a] MWSS
1	Tyrosine kinase-1	772307	1.990	−1.70	−2.21	−1.93
2	Serine/threonine kinase-1	403186	1.251	−1.90	−2.17	−2.15
3	Growth factor receptor	393930	1.365	−1.86	−2.41	−2.08
4	Serine/threonine kinase-2	814655	1.863	−1.73	−1.78	−1.41
5	Transcription factor	922084	0.889	−2.05	−2.49	−1.86
6	Heat shock protein	596679	0.359	−2.44	−2.44	−2.50
7	Metallocarboxypeptidase	128196	0.923	−2.03	−2.75	−2.52
8	Tyrosine kinase-2	217451	2.103	−1.68	−2.23	−2.10
9	Serine/threonine kinase-3	670430	0.362	−2.44	−2.53	−2.28
10	Serine/threonine kinase-4	941040	0.976	−2.01	−2.52	−2.13
11	Serine/threonine kinase-5	52422	5.700	−1.24	−1.80	−1.63
12	Tyrosine kinase-3	633398	2.164	−1.66	−2.02	−1.88

Predicted hit rates for Model 3 and MWSS are provided in last two columns

[a] Predicted hit rates with respective models

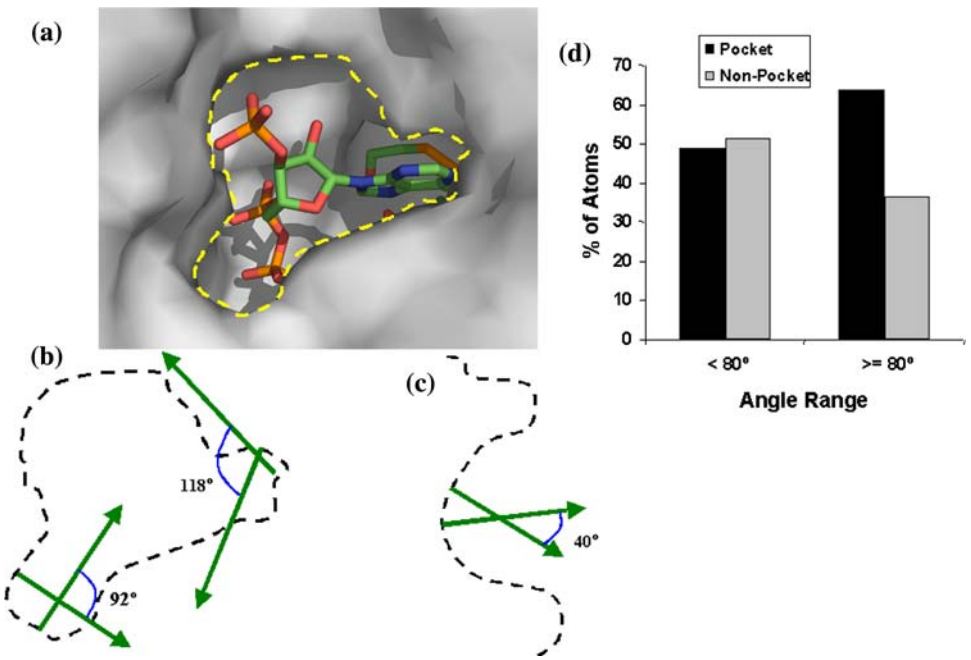

Fig. 2 The implementation of roughness parameter in the calculations **a** The ligand binding pocket features steep variations in surface as depicted by a 2D contour (*dashed lines*). **b** The variations of the surface can be quantified through surface normals (*green arrows*); patches of surface where the slope varies drastically (angles subtended by vicinity normals in excess of 80°) are common in ligand binding pockets. **c** A 2D contour of typical molecular surface cavity that may not bind to ligands; here normals most often display very acute angles with their neighbors representative of shallowness. **d** Distribution of atoms of 22 test target proteins over pocket and non-pocket area; the algorithm presented is able to distinguish between these classes of atoms merely based on angles subtended by normals with their neighborhood counterparts in the pocket and non-pocket regions

properties like—site points (enclosed grid points), exposure, enclosure, contact, hydrophobicity, hydrophilicity, ratio of phobic to philic (balance) and donor/acceptor points of hydrophilic regions. All the solvent exposed atoms of the target protein within 6 Å of the site points constituted the binding pocket region. Based on the size of the site points and molecular surface area of the binding pocket region (calculated using the *Molecular Surface* programme [23]), the shape factor was deduced as their ratio (volume/surface area). The number of charged residues, net formal charges of these residues as well as the partial charges of the exposed atoms present at the binding

site were considered as parameters. The partial charges of atoms were derived from AMBER force-field [24] using the templates given for standard amino acid residues. Since parameters such as principal axes and roughness are not part of most standard commercial packages of protein modeling, we devised procedures to calculate them. Once the set of atoms lining a binding pocket is obtained, the principal axes were computed as three mutually perpendicular longest axes that could be fitted within the volume of the binding pocket under consideration. Roughness is the extent of variations in the surface curvatures and is a measure of how rugged a surface patch is with respect to its immediate vicinity. For the computational purposes, we treated roughness as a sum of angular contributions of the normals of the surface points with respect to those of the neighboring vicinity of 3 Å. The larger angles reflect the high variability of surface slopes, which could contribute to a fine tuned specificity and affinity. When applied, our algorithm for roughness could differentiate between the pocket and the non-pocket residues (Fig. 2) in 78% cases of training dataset proteins in terms of pocket roughness being decisively greater than non-pocket roughness. In all we computed 28 molecular parameters along with their logarithmic values (with the exception of partial charge, net charges and *SiteMap* parameters). The complete set of parameters considered for the statistical regression exercise is tabulated in the Table 3.

Table 3 List of 28 structural parameters calculated for the training dataset proteins for characterization of their binding cavities and used for predictive model generation

S. no	Parameter	Abbreviation used
1	Molecular surface area	*msa*
2	Apolar molecular surface area	*amsa*
3	Polar molecular surface area	*pmsa*
4	Main chain molecular surface area	*main*
5	Side chain molecular surface area	*side*
6	Contact surface area	*csa*
7	Polar contact surface area	*pcsa*
8	Apolar contact surface area	*acsa*
9	Solvent accessible surface area	*sasa*
10	Polar solvent accessible surface area	*psasa*
11	Apolar solvent accessible surface area	*asasa*
12	Site score	*sitescore*
13	Size	*size*
14	Hydrophobiciy	*phobic*
15	Hydrophilicity	*philic*
16	Donor/acceptor	*Don/accpt*
17	Contact	*contact*
18	Exposure	*exposure*
19	Enclosure	*enclosure*
20	Balance	*balance*
21	Shape factor	*shape*
22	Roughness	*rough*
23	First principal axis	*fpa*
24	Second principal axis	*spa*
25	Third principal axis	*tpa*
26	Net formal charge	*netChrg*
27	Net partial charge	*parChrg*
28	Number of charged residues	*chrgRes*

Model generation

The above listed structural parameters were calculated for all 22 targets in our HTS training dataset. By the dictate of a standard regression protocol of data points to parameters ratio (5:1), this can support models with a maximum of four descriptors (structural parameters). We set out to determine the best set of four parameters (independent variables) that would contribute significantly to the logarithm of hit rate (dependent variable), by means of sequential multiple regression analysis. The regression analysis was conducted by using all theoretically possible combinations of four independent variables systematically to get the best set using the programme VALSTAT [25]. A couple of statistically significant models were considered on the basis of observed squared correlation coefficient (R^2), adjusted squared correlation coefficient (R^2_{adj}), standard error of estimate (*SEE*) and the sequential Fischer test (F). Internal predictivity of the models was ascertained with the help of constraints like cross-validated squared correlation coefficient (q^2), obtained by 'leave-one-out (*LOO*)' method. In this method, $N-1$ data points were used to build models and hit rate for Nth data point was predicted where N is the total number of data points ($N = 22$). The q^2 is defined as:

$$q^2 = 1 - \sum \left(Y_{pred} - Y_{act}\right)^2 \Big/ \sum \left(Y_{act} - Y_{mean}\right)^2$$

where, Y_{pred}, Y_{act}, and Y_{mean} are predicted, actual and mean values of the logarithm of hit rate, respectively. The external predictive power of the models has been analyzed with the help of test sets using predictive correlation coefficient (R^2_{pred}).

Results

Model analysis

In the first cycle we generated 10 models and found that the 3rd model (Model 3) was statistically significant. In addition, another model was generated without *SiteScore* (MWSS) parameter as *SiteScore* contained some of the

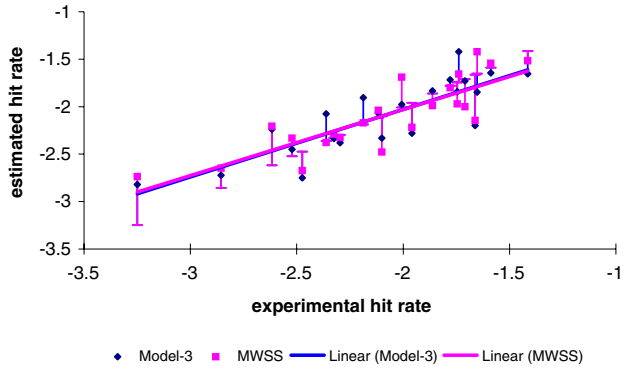

Fig. 3 Correlation between experimental HTS hit rates and estimated hit rates for 22 proteins of training dataset with Model 3 and MWSS. Where R^2 values are 0.71 and 0.70, respectively. *Error bars* shown in the plot represent residual values for each estimated hit rate

Table 4 Y randomization results for Model-3 and MWSS

Parameters/models	Model-3	Model-MWSS
Randomized R^2_{max}	0.607	0.613
Randomized R^2_{mean}	0.189	0.195
Randomized standard deviation (S_{rand})	0.115	0.119

Table 5 Significant parameters contributive to prediction of hit rates in Model 3 and MWSS with their regression coefficients and statistical parameters describing the robustness of the three models

Parameters	Model 3	MWSS
Structural		
amsa	0.0029 (0.0004)	–
pmsa	−0.0023 (0.0007)	–
log_pmsa	–	−4.9236 (1.8550)
fpa	−0.0576 (0.0288)	–
log_main	–	1.5683 (1.3566)
sitescore	−1.9334 (0.5568)	–
size	–	0.0120 (0.0046)
shape	–	−22.2752 (10.3839)
Constant	0.6148	7.4230
Statistical		
R^2	0.712	0.696
R^2_{adj}	0.645	0.624
SEE	0.272	0.279
F value	10.521	9.729
LOO q2	0.506	0.520

structural parameters implicitly. Model 3 and MWSS have correlation coefficients (R) of 0.844 and 0.834, respectively, which accounts for more than 64.4 and 62.4% of the explained variance in the activity, respectively, calculated as $R^2_{adj} = R^2(1 - 1/F)$. The data showed an overall internal statistical significance level better than 99% as the calculated variance ratio, i.e., Fischer value (F) exceeded the tabulated $F_{(4,17\ \alpha\ 0.01)} = 4.67$ suggestive of applicability of models in 99 out of 100 instances. Low value of *SEE* suggests a high degree of confidence in the analysis. The cross-validated squared correlation coefficients (q^2) of Model 3 and MWSS were 0.505 and 0.520, respectively, suggesting a moderate internal consistency as well as predictive ability of the HTS hit rate (Fig. 3). The chance of fortuitous correlation was checked by Y randomization (randomization of response) approach [26] for 1,000 cycles. Mean R^2 values for the selected models (Model-3 and MWSS) are less than 0.2 suggesting that the results are not based on chance correlation (Table 4). The coefficients of main descriptors influencing the predicted hit rate and statistical regression parameters for both models are summarized in Table 5. The data within the parentheses with the coefficient value is the standard deviation associated with it. We notice that with *SiteScore* included, three parameters that could play a role in determining the hit rates are apolar molecular surface area (*amsa*), polar molecular surface area (*pmsa*) and first principal axis (*fpa*) (Model 3). Without the inclusion of *SiteScore* (MWSS model) the dominant parameters are the *size* (volume), logarithm of *pmsa*, logarithm of *main* and *shape* performing very close to the Model 3 in terms of regression fitting. With such acceptable values of regression parameters, the stage was set to test the validity of our models (Model 3 and MWSS) on a different test dataset.

Table 6 Performance of Model 3 and MWSS on in-house HTS and Hajduk et al. training datasets

Statistical models	Correlation coefficients, R^a_{pred} (R^2_{pred} values in square parenthesis)	
	In-house test dataset with HTS HR[b] (12 proteins)	Hajduk et al. training dataset with NMR screening HR[b] (11 proteins)
Model 3	0.74 [0.55]	0.50 (with one false negative[c]) [0.25]
MWSS	0.70 [0.49]	0.58 [0.34]

[a] Predicted hit rates versus experimental hit rates

[b] Hit rate

[c] High affinity drug like ligand present but cannot predict

Predictions with HTS test dataset

For the 12 targets in the test dataset, the structural coordinates file information, the experimental hit rates, as well as the hit rates predicted by both models in our exercise

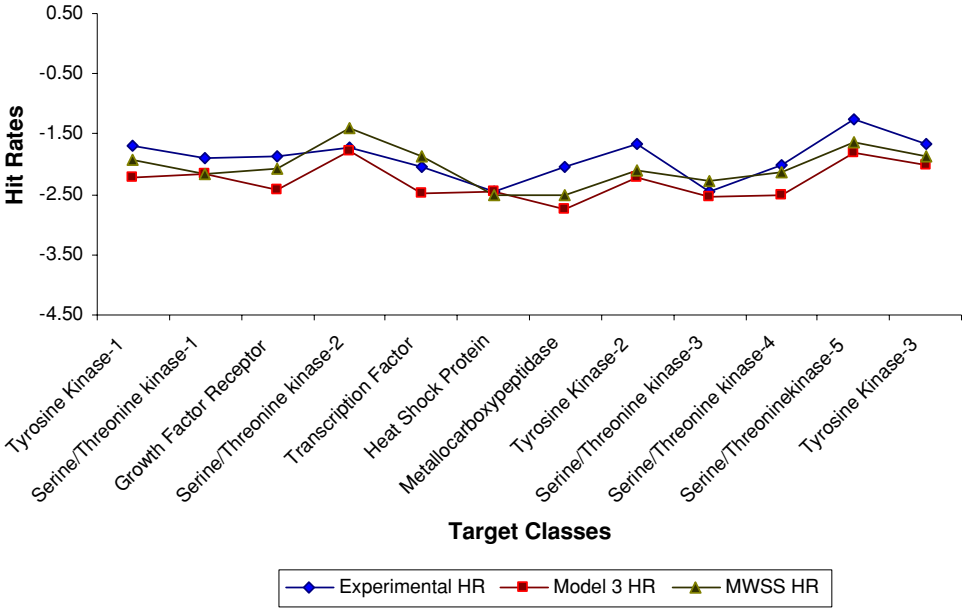

Fig. 4 Predicted hit rates derived from Model 3 (*squares*) and MWSS (*triangles*) are aligned with actual HTS hit rates (*diamonds*) for various targets in test dataset. The correlation coefficients (R_{pred}) are 0.74 and 0.70, respectively

have been tabulated in Table 2. Predicted hit rates were calculated using the equation below:

$$HR_{pred} = \sum_{i=1}^{i=4} a_i X_i$$

where, HR_{pred} is the predicted Hit rate, a is the ith parameter coefficient and X is the parameter value.

The correlation coefficients (R_{pred}) along with R^2_{pred} between experimental and the predicted hit rates are given in Table 6. It is immediately apparent from this data that both the models have a decent performance with a correlation coefficient of 0.7 or more. We examined the degree of agreement of the models with the HTS test data set. It can be seen that for most targets (e.g. serine-threonine kinase-1 and 2 and binding protein) there is a high degree of qualitative agreement between the profiles of actual hit rates and predicted hit rates (Fig. 4). The predictive ability of Model 3 is exemplified by the quality of agreement at both ends of hit rates spectrum; the targets within top two of three (serine-threonine kinase-5 and tyrosine kinase-3) and bottom five (transcription factor, metallocarboxypeptidase, heat shock protein, serine-threonine kinase-3 and 4) hit rates of HTS test data are identified by this model in their respective categories. To understand the extent of influence of the size of the screening library on the predictability of the models, we looked at the correlation between the sizes of the libraries used for screening with difference between experimental and predicted hit rates. While no definitive inference could be drawn for Model 3, the correlation coefficient of 0.77 for MWSS ascertained that larger size has a better statistical significance.

Discussion

The definition and data pertaining to the proteins that bind to high-affinity ligands (druggable proteins) is relatively on a firmer ground. However, there is a considerably lesser clarity on what may be "experimentally proven non-druggable proteins" which may be driven by the context. In the pharmaceutical industry a protein with a very low or no hit rate with HTS against their corporate library may be a non-druggable protein. Thus the compilation of information on non-druggable proteins across industry and academia is indeed a difficult proposition. This is mainly due to the limitations of systematic databases that combine disparate data that may reside within a company warehouse, heterogeneity in formats where this information may be available or inadequate compulsions to publish the so-called negative results [27]. Hence the models developed to predict the HTS hit rates can be applied along with other essential parameters (like biological significance of the protein etc.) in order to determine whether a protein-binding site can be druggable or decoy.

Extraction of proprietary HTS subset

We recognize that although datasets comprising Kd values are very readily accessible, Kd is a measure of binding affinity of a *given ligand* and not of a protein-binding pocket. Kd dataset's limitation of dependence on the available data—*viz*., the results could vary in the light of experiments with new set of ligands, makes it a weak index to pursue on protein druggability. The hit rates of HTS

experiments are a better measure of protein's ability to bind to a diverse, large number of small molecules bearing drug-like properties. Beyond reasons such as choice of targets, compound collections etc. that could influence the outcome of HTS campaigns, the rationale behind adhering to HTS data as the source to train models in our investigation rests mainly on access to experimental data displaying a range of binding affinity of small molecules with the targets. In the Discovery context, the outcome is mainly assessed based on how many of the hits identified translated into potent leads and so on, whereas we would consider HTS data that could relate modulation of target enzyme character by virtue of a small molecule binding to its binding pocket in the context of an assay. It has to be emphasized here that binding is taken in the same sense as binding to the main pocket whose structural descriptors are analyzed. CCT is a convenient way to concentrate on target features rather than on variations due to different set of compounds used for HTS. This way the hit rates of CCT represent true variation among the binding sites in their ability to bind to small molecules. The CCT compounds belonged to 14,891 clusters (9,971 singletons) when a *Tanimoto*[1] cut-off of 0.3 was applied reflecting a considerable chemical diversity within this collection.

While applying our models on various test datasets we looked for (a) correlation of our predicted hit rate versus the experimental hit rate corresponding to the dataset and (b) occurrence of false negative in our prediction. In any predictive exercise false negatives should be viewed more seriously than the instances of false positives. Thus minimal or no false negatives should be indicative of a robust model.

Performance of the models with in-house and external dataset

Predictive ability of these two models was validated on in-house HTS test dataset. This dataset contains 12 proteins that were not used in training dataset and for which experimental HTS hit rate values were available. Even though the training data set was not derived with any particular target class in focus, we find that the test data set in which kinases as a class significantly represented, the correlation coefficient between (R_{pred}) the experimental and predicted hit rates is 0.77. R^2_{pred} for Model 3 and MWSS are 0.549 and 0.501, respectively. These values suggest acceptable predictive power of the generated models.

The dataset used by Hajduk and co-workers [5] contained 23 proteins of which we had the access to the crystallographic ligand bound structures of 11 proteins; these were added to the validation dataset. In the application to Hajduk training dataset we notice that among 11 target proteins, 5 are known to bind to high affinity drug like molecules (highly druggable). Our model MWSS, in fact is sound in its prediction (Table 6) with a correlation of 0.58 between the hit rates. Model 3 is not far behind with only one instance (out of 11) of high affinity drug like ligand present but not predicted.

Among 28 parameters and their logarithmic values that we analyzed with the view to determine their individual and collective impact on the target hit rates, by virtue of being primary constituents of either Model 3 or MWSS, seven descriptors *viz*., size, polar molecular surface area, logarithm of main chain molecular surface area, apolar molecular surface area, first principal axis, *SiteScore* and shape factor, significantly bear upon the ability of a target binding site to attract drug-like molecules as elucidated by this model's consistency in obtaining acceptable statistical indices. A few aspects of such an outcome are noteworthy. For instance, a critical size or volume is needed for steering in and maneuvering the ligand molecule into the binding site. On the other hand larger volume does not necessarily bring about effective affinity. Just as ligand efficiency [28, 29] (which is free energy of binding per non-hydrogen atom) is a key consideration in optimization of interaction from small molecules perspective, compactness (surface area per unit volume) serves as a structural analogue from the protein binding site point of view. Large negative coefficient for *shape* (inverse of compactness) translates to positive impact of compactness. Larger surface area or interaction points per unit volume drive pocket efficiency. The coefficient for logarithm of *pmsa* is negative (in MWSS) or non-contributive (Model 3) indicating that the polar interactions may not be vital in a binding event. The fact that druggability predicted by Hajduk and coworkers was dominated by pocket shape and hydrophobicity rather than polar interactions prompting them to suggest that the primary role for charged interactions may be in imparting specificity rather than potency in the binding process. Similar interpretation could be accorded to the parameter roughness, which is able to distinguish pocket atoms from non-pocket atoms on its own, but did not show up in the top four descriptors in both models. It is our view that rather than attempting to extract finer interpretations of the structural descriptors presented here, holistic view of models generated would be of further physical/biological relevance. Recognizing that such mathematically derived models can be significantly influenced by the datasets and the experiments behind them, we benchmarked their performance by using the dataset derived from very different screening protocols (HTS and NMR fragment screening) and are comforted by their broader applicability.

[1] Tanimoto TT, IBM Internal Report, November 17, 1957.

Summary

Target identification and validation are vital steps that impact further cascade of investigations in rational target based drug discovery processes. The established procedures that determine the essentiality and selectivity help determine the candidacy of a potential target. From the point of view of committing resource on a target for downstream transitions, it is imperative to augment experimental strategies with due and complementary computational and knowledge-based efforts. In its scope our work is a step towards ensuring that a set of targets under consideration can be prioritized given the knowledge of their three dimensional structural information. It includes exercises of employing diverse datasets in an attempt to identify coefficients (weights) for various structural and physicochemical parameters through different regression models. We use the results of training of the model based on in-house HTS data to qualify the models (Model 3 and MWSS) to demonstrate their applicability over other datasets including a different in-house HTS dataset where they yielded very agreeable R^2 value and R^2_{pred} value of around 0.7 and 0.5, respectively. The consistency in performance of these models over various datasets indicates that the parameters size, polar molecular surface area, main chain molecular surface area, shape factor, apolar molecular surface area, first principal axis and *SiteScore* stand out of 28 parameters considered in terms of their contribution in determining the structural druggability of a target protein. This conclusion seems to be somewhat independent of the choice of dataset i.e., experiments used to obtain them (in both training and testing). However the complex nature of the relationship between the structural descriptors and druggability (or affinity to bind) suggests that before any prioritization of targets could be done, varying scenarios of regression need to be considered. Hence we present both the models (Model 3 and MWSS) rather than being too prescriptive with a single model. The targets that are predicted to possess higher hit rates by both models could be the basis to label them as candidates to give significant hits in HTS campaigns. We acknowledge that there is still some room for the parameters and the specific weights derived for them to be influenced by the nature of the screening databases or other derived descriptors and hope that the viewpoints of this exercise could be of use to a wider scientific community, for further developments in the area of computational assessment of protein druggability in order to seek targets with prospect.

Acknowledgments We would like to thank Dr. Stefan Schmitt (SS) for offering valuable suggestions during the course of this project. We are also grateful to SS, Drs. Bheemarao Ugarkar, Manoranjan Panda and Raghuram Tangirala for their comments on the manuscript.

References

1. Betz UA (2005) How many genomic targets can a portfolio afford? Drug Discov Today 10(15):1057–1063. doi:10.1016/S1359-6446(05)03498-7
2. Hopkins AL, Groom CR (2002) The druggable genome. Nat Rev Drug Discov 1(9):727–730. doi:10.1038/nrd892
3. Drews J (2006) Whats's in a number? Nat Rev Drug Discov 5:975. doi:10.1038/nrd2205
4. Zambrowicz BP, Sands AT (2003) Knockouts model the 100 best-selling drugs–will they model the next 100? Nat Rev Drug Discov 2(1):38–51. doi:10.1038/nrd987
5. Hajduk PJ, Huth JR, Fesik SW (2005) Druggability indices for protein targets derived from NMR-based screening data. J Med Chem 48(7):2518–2525. doi:10.1021/jm049131r
6. Han LY, Zheng CJ, Xie B, Jia J, Ma XH, Zhu F, Lin HH, Chen X, Chen YZ (2007) Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. Drug Discov Today 12(7–8):304–313. doi:10.1016/j.drudis.2007.02.015
7. Nayal M, Honig B (2006) On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. Proteins 63(4):892–906. doi:10.1002/prot.20897
8. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, Salzberg AC, Huang ES (2007) Structure-based maximal affinity model redicts small-molecule druggability. Nat Biotechnol 25(1):71–75. doi:10.1038/nbt1273
9. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are the there? Nature 5(12):993–996. doi:10.1038/nrd2199
10. Lipinski CA (2000) Drug-like properties and the causes of poor solubility and poor permeability. J Pharmacol Toxicol Methods 44(1):235–249. doi:10.1016/S1056-8719(00)00107-6
11. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46(1–3):3–26. doi:10.1016/S0169-409X(00)00129-0
12. Kuntz ID, Chen K, Sharp KA, Kollman PA (1999) The maximal affinity of ligands. Proc Natl Acad Sci USA 96(18):9997–10002. doi:10.1073/pnas.96.18.9997
13. Bajorath J (2002) Integration of virtual and high-throughput screening. Nat Rev Drug Discov 1(11):882–894. doi:10.1038/nrd941
14. Davies JW, Glick M, Jenkins JL (2006) Streamlining lead discovery by aligning in silico and high-throughput screening. Curr Opin Chem Biol 10(4):343–351. doi:10.1016/j.cbpa.2006.06.022
15. Pereira DA, Williams JA (2007) Origin and evolution of high throughput screening. Br J Pharmacol 152(1):53–61. doi:10.1038/sj.bjp.0707373
16. Pellecchia M, Bertini I, Cowburn D, Dalvit C, Giralt E, Jahnke W, James TL, Homans SW, Kessler H, Luchinat C, Meyer B, Oschkinat H, Peng J, Schwalbe H, Siegal G (2008) Perspectives on NMR in drug discovery: a technique comes of age. Nat Rev Drug Discov 7:738–745. doi:10.1038/nrd2606
17. Puvanendrampillai D, Mitchell JB (2003) L/D Protein ligand database (PLD): additional understanding of the nature and specificity of protein-ligand complexes. Bioinformatics 19(14):1856–1857. doi:10.1093/bioinformatics/btg243

18. Kumar MD, Gromiha MM (2006) Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein–ligand complexes. Nucleic Acids Res 34(Database issue): 195–198. doi:10.1093/nar/gkj017
19. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res 35(Database issue):198–201. doi:10.1093/nar/gkl999
20. Perola E, Walters WP, Charifson PS (2004) A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance Proteins. Proteins 56(2):235–249. doi:10.1002/prot.20088
21. Böhm HJ (1998) Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. J Comput Aided Mol Des 12(4):309–323. doi:10.1023/A:1007999920146
22. Schrödinger SiteMap Fast, accurate and practical binding site identification. 8.0. (2008) New York, NY, Schrödinger, LLC. 2005. Ref Type: Computer Program
23. Connolly ML (1993) The molecular surface package. J Mol Graph 11(2):139–141. doi:10.1016/0263-7855(93)87010-3
24. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ (2005) The Amber biomolecular simulation programs. J Comput Chem 26(16):1668–1688. doi:10.1002/jcc.20290
25. Gupta AK, Babu MA, Kaskhedikar SG (2004) VALSTAT : validation program for quantitative structure activity relationship studies. Indian J Pharm Sci 66(4):396–402
26. Wold S, Eriksson L (1995) Statistical validation of QSAR results. In: van de Waterbeemd H (ed) Chemometrics methods in molecular design. VCH, Weinheim, pp 309–318
27. Veretnik S, Fink JL, Bourne PE (2008) Computational biology resources lack persistence and usability. PLOS Comput Biol 4(7):e1000136. doi:10.1371/journal.pcbi.1000136
28. Abad-Zapatero CMJT (2005) Ligand efficiency indices as guideposts for drug discovery. Drug Discov Today 10(7):464–469. doi:10.1016/S1359-6446(05)03386-6
29. Hopkins AL, Groom CR, Alex A (2004) Ligand efficiency: a useful metric for lead selection. Drug Discov Today 9(10):430–431. doi:10.1016/S1359-6446(04)03069-7