Molecular Diversity (2006)
DOI: 10.1007/s11030-006-9035-3

Perspective

Interrogating the druggable genome with structural informatics

Kevin Hambly*, Joseph Danzer, Steven Muskal & Derek A. Debe
Eidogen-Sertanty, Inc., 9381 Judicial Dr., Suite 200, San Diego, CA 92121, USA
*(*Author for correspondence, E-mail: khambly@eidogen-sertanty.com, Tel.: +858-964-2092, Fax: +775-822-1721)*

Received 18 December 2005; Accepted 13 April 2006

Key words: structure–based drug design, druggable genome, protein structures, ligand binding sites, drug targets, structural genomics

Summary

Structural genomics projects are producing protein structure data at an unprecedented rate. In this paper, we present the Target Informatics Platform (TIP), a novel structural informatics approach for amplifying the rapidly expanding body of experimental protein structure information to enhance the discovery and optimization of small molecule protein modulators on a genomic scale. In TIP, existing experimental structure information is augmented using a homology modeling approach, and binding sites across multiple target families are compared using a clique detection algorithm. We report here a detailed analysis of the structural coverage for the set of druggable human targets, highlighting drug target families where the level of structural knowledge is currently quite high, as well as those areas where structural knowledge is sparse. Furthermore, we demonstrate the utility of TIP's intra- and inter-family binding site similarity analysis using a series of retrospective case studies. Our analysis underscores the utility of a structural informatics infrastructure for extracting drug discovery-relevant information from structural data, aiding researchers in the identification of lead discovery and optimization opportunities as well as potential "off-target" liabilities.

Introduction

The completion of the human genome in 2001 revealed that the total number of human genes is somewhere between 30,000 and 40,000 [1]. Recent analysis, however, has suggested that the number of potential drug targets coded by the human genome may in fact be much smaller than originally speculated, with all currently marketed drugs being directed at only 120 unique targets [2]. Homology-based extrapolation of this number to the entire genome indicates that the total number of potentially druggable targets may range from 3000 to 5000, with only a subset of these targets expected to be directly linked to a disease state [3].

While novel, druggable, and clinically relevant targets may represent a limited subset of the genome, these targets are commonly members of larger protein families whose constituents share many of the same sequence, structure, and binding site characteristics as the primary druggable target. Complicating matters, these similar targets may not be relevant to the particular disease state due to differing tissue distribution, expression levels, or regulation under different physiological conditions. Hence, even though the absolute number of druggable, disease-relevant targets is limited, the number of undesirable "off-targets" for any given small molecule modulator can be quite large, supporting the well established need for technologies capable of uncovering potential selectivity and cross-reactivity issues as early in the drug development process as possible.

Following the completion of the human genome, structural genomics emerged as a major research initiative charged with providing structural and functional information for all human proteins. Two of the most well-funded structural genomics efforts are the Protein Structure Initiative (PSI) [4], which is focused primarily on solving one structure for every major protein fold family, and the recently launched Structural Genomics Consortium (SGC) [5], which is focused almost entirely on solving the structures of human proteins. Given its focus on human proteins, the SGC project is expected to significantly expand the experimental structural information available for the "druggable" human genome.

During and after the elucidation of the human genome, the field of bioinformatics matured very quickly to address the need for the useful organization and analysis of the emerging landslide of gene sequence data. A computational community of similar size and resolve has yet to emerge for the organization and analysis of the landslide of structural data produced by structural genomics. Indeed, the well known structural family classification resources such as SCOP [6], CATH [7],

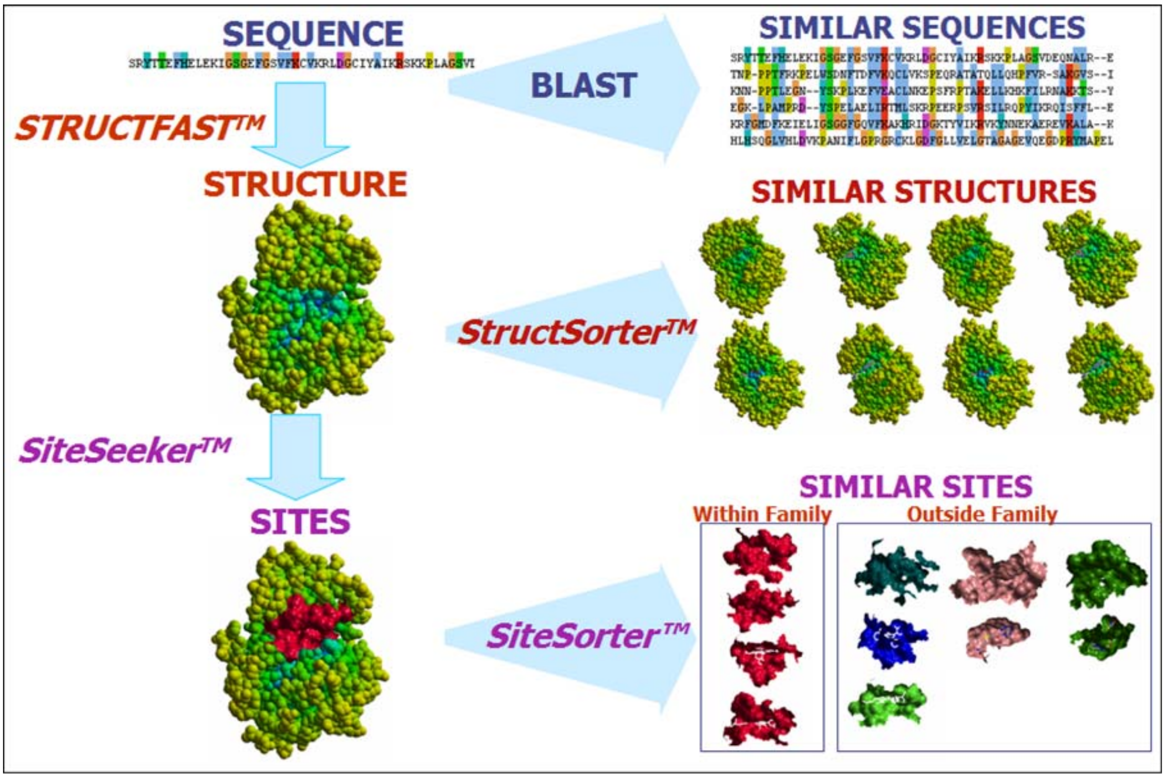
SEQUENCE
SRYTTEFHELEKIGSGEFGSVFKCVKRLDGCIYAIKRSKKPLAGSVI
BLAST
SIMILAR SEQUENCES
SRYTTEFHELEKIGSGEFGSVFKCVKRLDGCIYAIKRSKKPLAGSVDEQNALR--E
TNP-PPTFRKPELWSDNFTDFVKQCLVKSPEQRATATQLLQHPFVR-SAKGVS--I
KNN-PPTLEGN--YSKPLKEFVEACLNKEPSFRPTAKELLKHKFILRNAKKTS--Y
EGK-LPAMPRD--YSPELAELINTMLSKRPEERPSVRSILRQPYIKRQISFFL--E
KRFGMDFKEIELIGSGGFGQVFKAKHRIDGKTYVIKRVKYNNEKAEREVKALA--K
HLHSQGLVHLDVKPANIFLGPRGRCKLGDFGLLVELGTAGAGEVQEGDPRYMAPEL
STRUCTFAST™
STRUCTURE
SIMILAR STRUCTURES
StructSorter™
SiteSeeker™
SITES
SIMILAR SITES
Within Family
Outside Family
SiteSorter™

Figure 1. Overview of the algorithms applied within the target informatics platform. These algorithms are used to amplify the initial information contained in the TIP database: the protein sequences from the International Protein Index database (IPI) [11], and the protein structures from the Protein Data Bank (PDB) [12]. The algorithm engine in TIP proceeds as follows: First, the sequence similarity relationships are calculated using the BLAST algorithm [13]. Next, structures are determined for each sequence using the homology modeling algorithm, STRUCTFAST [14]. After the structures have been determined, their associated small molecule binding and protein-protein interaction sites are determined using the site finding algorithm, SiteSeeker [15]. After this step, the similarity relationships between each of the structures are calculated using the structure alignment algorithm StructSorter. Finally, the binding site similarity relationships are calculated using a weighted clique detection algorithm called SiteSorter [16]. It has been independently reported elsewhere that clique detection algorithms are capable of determining similarities between small molecule binding and protein-protein interaction sites on targets that do not share any sequence or structure homology [17]. Since structure determination, structure alignment, and site alignment require significantly more computation time than sequence alignment via BLAST, a database has been integrated into TIP to store the results of these calculations and automatically initiate new calculations when new experimental structure data is uploaded. In a separate publication, we have reported on a clustering methodology that allows us to continuously update and maintain the TIP database of structural alignments in a computationally efficient manner[18]. These calculations required approximately 3 months to complete on a 128-node Linux cluster (3 GHz processors) for the ~30,000 structures in the PDB and the ~25,000 human target sequences in TIP. Currently, we are calculating other drug discovery relevant proteomes, such as mouse and rat, and various pathogenic species.

and FSSP [8], Gene3D [9], and VAST [10], have yet to be extended to address the drug discovery relevant problem of target binding site similarity and cross-reactivity.

The incredible progress made by the experimental protein structure community in this decade establishes two important challenges to the computational community. First, since the various structural genomics projects are not projected to complete the human structural proteome within the next two decades, there is a need for computational approaches that are capable of amplifying the existing structural data to broaden the current structural coverage of druggable target space. Second, since the probability of having multiple structures per target family has increased significantly, there is a need for new informatics approaches capable of leveraging structural data in a holistic manner to enhance the discovery and optimization of selective small molecule protein modulators on a genomic scale.

In this paper, we report a novel structural informatics approach to storing, organizing, and amplifying the growing body of experimental protein structure data. Since the precise details of the database architecture and algorithmic methods utilized in our approach are beyond the scope of this publication and will be published elsewhere, this paper will outline the general framework of our approach and report on its capacity to address the two computational challenges outlined, specifically addressing how our approach has been applied for annotating the druggable human genome with structural information.

Methods

The Target Informatics Platform (TIP) consists of a fully automated computational approach for determining protein

# Human Druggable Target (DT) PFAM Domains	**161**	
# Unique DT domain sequences in IPI and PDB *(clustered at 95% ID)*	**5930**	
# DT sequences with solved structure in PDB	**740**	**12.5%**
#DT sequences which can be modeled	**4011**	**67.6%**
...at >70% ID to PDB template	*747*	
...at between 50% and 70% ID	*396*	
...at between 30% and 50% ID	*1200*	
...at between 20% and 30% ID	*818*	
...at <20% ID	*850*	
# DT sequences with no PDB or Model	**1179**	**19.9%**

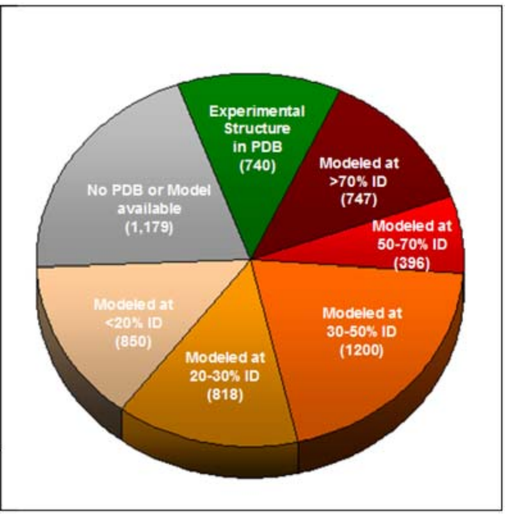

Figure 2. Structural annotation statistics for 5,930 druggable target PFAM domain sequences. The total number of unique Drug Target (DT) PFAM domain sequences corresponds to all of the unique sequence regions found in human sequences from the IPI dataset, as well as sequences from PDB derived from human and the related mammalian species *Bos taurus* (cow), *Mus musculus* (sheep), *Oryctolagus cuniculus* (rabbit), *Rattus norvegicus* (rat), and *Sus scrofa* (pig). The percent identity ranges shown for modeled sequences correspond to the percent sequence identity to the PDB template used for modeling.

structures and binding sites and their respective similarities, as well as a database to store the calculation results so that they are available in the future without the need for additional calculations. Figure 1 shows an overview of the algorithms that are applied within TIP.

Results and discussion

While the performance of the STRUCTFAST, StructSorter, SiteSeeker, and SiteSorter algorithms are being considered in separate publications, we report here on the potential of a structural informatics platform such as TIP to address the two important computational challenges outlined in the introduction. First, we will discuss in detail TIP's structural coverage for druggable human targets. Second, we discuss several examples of interesting and well-established cross-reactivities between druggable target families that can be revealed via the synergistic application of the STRUCTFAST, SiteSeeker, and SiteSorter algorithms.

Structural coverage for druggable target space

To define the space of known and potential druggable targets in the human genome, we used a sequence-domain based approach for annotation and retrieval of protein sequences containing known "druggable domains". As the starting point for this analysis, we used a set of 125 druggable Interpro [19] domains that were described by Hopkins et al. [2] as functional domains which have been shown to bind compounds obeying the Lipinski Rule-of-5 criteria for drug-likeness [20]. Using the database cross-referencing feature of Interpro, we mapped each of these 125 Interpro domains to a unique domain from the PFAM Protein Families database [21]. This list of PFAM domains was supplemented with PFAM domains from a manually curated internal database of potentially druggable targets, to yield a total of 161 unique PFAM domains associated with druggable targets. For all of the analysis performed here, the domain sequences described correspond to *only* the PFAM domain regions, not the entire protein sequence. There are in fact numerous cases where multiple distinct "druggable" PFAM domains exist within a single protein sequence, so all of these regions are counted as unique domains in this work.

To extract a set of human sequences containing these druggable PFAM domains we used RPS-BLAST [22] to query a set of human sequences derived from the International Protein Index (IPI) and PDB against the set of PFAM domain profiles. The final list of IPI and PDB sequences was clustered at 95% ID to yield a total of 5,930 unique druggable domain sequences.

Querying TIP with the 5,930 druggable target domain sequences yielded 740 sequences associated with experimental structures in the PDB, 4,011 sequences associated with models built by the STRUCTFAST comparative modeling algorithm incorporated into TIP, and 1,179 sequences which were not associated with any PDB structure nor any modeled structure (Figure 2).

As shown in Figure 2, of the 4,011 druggable sequences with homology models available, there is a great deal of variability in terms of the "resolution" with which these models were built, or in other words, the level of similarity to the PDB template used for modeling. While 747 sequences can be modeled very reliably with similarity levels of >70% ID to the template, over 40% (1668) of the sequences which can be modeled had less than 30% ID to the best PDB template, with nearly half of these (850) built at very remote similarity levels of less than 20% ID. To put this in context, in general

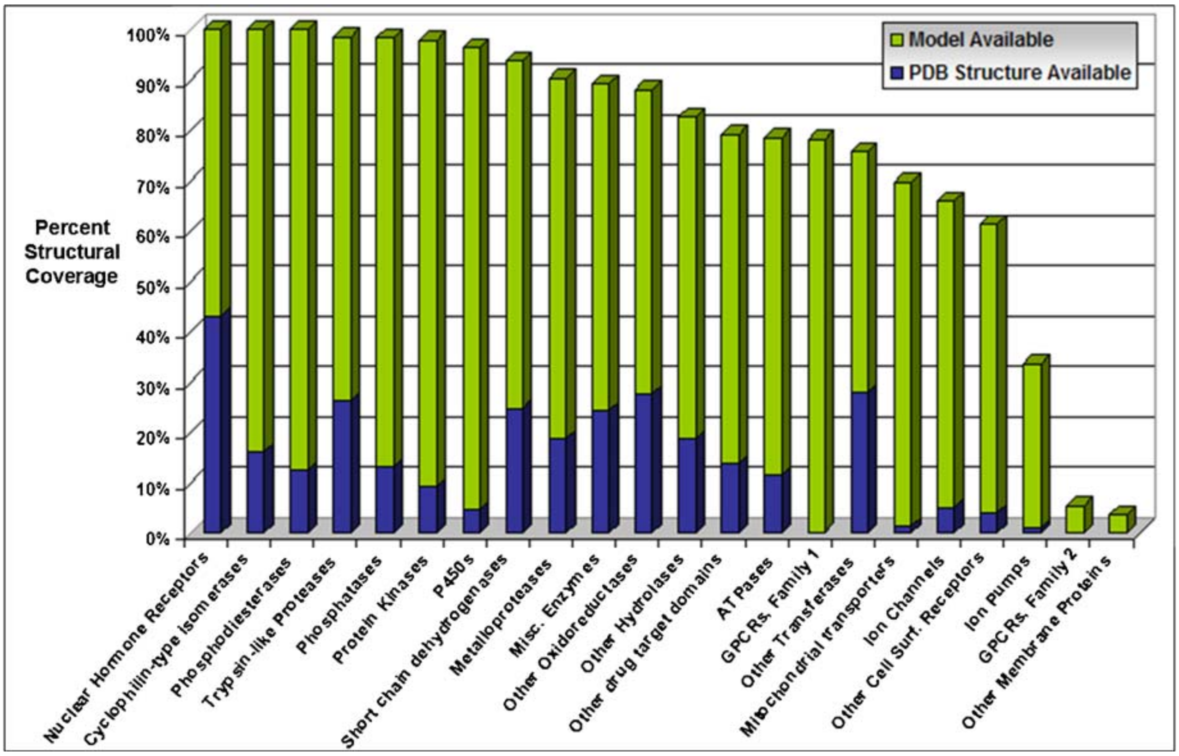

Figure 3. Druggable Target Family Percent Structural Coverage. Blue bars represent the percentage of members of a given target family that have experimental structures available in the PDB. Green bars indicate the percentage of targets which have homology models available in the TIP database.

it is accepted that models with greater than 50% sequence identity are usually of a high enough quality to have true utility in structure-based drug design and lead optimization based on a detailed understanding of ligand-binding interactions [23]. Models that have a sequence identity between 30% and 50% tend to be more confidently used for target druggability assessments, ligand binding site identification, and structure-based virtual screening. Sequence similarity below 30% is often regarded as the "twilight zone" of comparative modeling – where the confidence level in structural and/or functional annotations via homology inference tends to drop. Nevertheless, while the ligand binding sites of such remote homology models may not be resolved well enough for structure-based drug design, the overall features of the structure can still have significant utility for initial selectivity analyses and the design of mutagenesis experiments, for example.

Target family structural coverage

To get a better sense of the distribution of structural coverage by target family, we performed a classification of the druggable PFAM domains in order to cluster similar targets together and determine the structural coverage for each family. Table 1 shows the structural annotation available for the 22 target groupings that were used (numerous smaller target families are grouped together according to class). This analysis highlights that the overall number of druggable domain sequences which are represented in the PDB is still quite low at only 12.5%. Some target families, however, have significantly better coverage in the PDB than others, such as Nuclear Hormone Receptors at 43% (41 of 95 total sequences are associated with a PDB), and Trypsin-like proteases at 26% (56 of 214 sequences). The lowest level of experimental structural coverage is found in members of the membrane protein target families, which is not surprising due to the ongoing challenges posed by membrane-protein crystallography [24].

If we consider structural coverage to include comparative models as well as PDBs, then a much more complete structural landscape is observed, with 80% of all druggable domains having either a PDB structure or modeled structure. By combining comparative models and PDB structures, some gene families such as Nuclear Receptors, Phosphodiesterases, and Cyclophilin-type isomerases are found to in fact have 100% structural coverage, while larger families such as Protein Kinases, Trypsin-Like Proteases, and Phosphatases are not far behind at approximately 98% coverage (Figure 3). While many of these models are still not at the highest level of resolution, as illustrated in Figure 2, it is nevertheless encouraging to see that, with the exception of membrane-bound proteins, structural information for most members of major target families can now be obtained via comparative modeling.

Table 1. Statistics for target-family specific structural annotation of druggable PFAM domains. PFAM Domains have been grouped according to their classification as Enzymes, Membrane Proteins, Nuclear proteins, or Miscellaneous. PFAM IDs are listed for those target families with less than seven associated PFAM domains in the group. The number of sequences corresponds to the number of unique sequence domains within a single sequence. In cases where multiple distinct druggable PFAM domains are found within a single protein sequence, then both domains are counted.

PFAM ID(s)	Class	Family	Total # Sequences	#Seq. w/PDB	% PDB Coverage	#Seq. w/Model	% Model Coverage
PF00004, PF00122, PF02518, PF00689	Enzyme:hydrolase	ATPases	149	17	11%	100	67%
9 PFAM domains	Enzyme:Hydrolase	Metalloproteases	166	31	19%	119	72%
29 PFAM domains	Enzyme:Hydrolase	Misc. Hydrolases	572	105	18%	382	67%
PF00102, PF00782	Enzyme:Hydrolase	Phosphatases	130	17	13%	111	85%
PF00233	Enzyme:Hydrolase	Phosphodiesterases	48	6	13%	42	88%
PF00089	Enzyme:Hydrolase	Trypsin-like Proteases	214	56	26%	155	72%
PF00160	Enzyme:Isomerase	Cyclophilin-type isomerases	75	12	16%	63	84%
15 PFAM domains	Enzyme:Isomerase	Misc. other enzymes	149	36	24%	97	65%
23 PFAM domains	Enzyme:Oxidoreductase	Misc. oxidoreductases	343	94	27%	208	61%
PF00067	Enzyme:Oxidoreductase	P450s	110	5	5%	101	92%
PF00106	Enzyme:Oxidoreductase	Short chain dehydrogenases	97	24	25%	67	69%
28 PFAM domains	Enzyme:Transferase	Misc. transferases	222	62	28%	106	48%
PF00069	Enzyme:Transferase	Protein kinases	738	67	9%	656	89%
PF00001	Membrane:Cell surf. rec.	GPCRs, family 1	945	1	0%	738	78%
PF00002	Membrane:Cell surf. rec.	GPCRs, family 2	94	0	0%	5	5%
7 PFAM domains	Membrane:Cell surf. rec.	Other cell surf. receptors	129	5	4%	74	57%
PF02931, PF02932, PF01007, PF00654, PF00060, PF00858	Membrane:Transport prot.	Ion channels	138	7	5%	84	61%
PF00520, PF00209, PF00955, PF01758	Membrane: Transport prot.	Ion pumps	235	2	1%	77	33%
PF00153	Membrane: transport prot.	Mitochondrial transporters	76	1	1%	52	68%
PF01490, PF00664, PF00324	Membrane: transport prot.	Other membrane proteins	110	0	0%	4	4%
PF00104	Nuclear	Nuclear hormone receptors	95	41	43%	54	57%
21 PFAM domains	Misc.	Other drug target domains	1095	151	14%	716	65%
		Grand Total	5930	740 12.5%		4011 67.6%	

Table 2. Summary of TIP's binding site content for all available druggable target structures. The total number of structures corresponds to all unique PDB IDs for that family, plus the number of homology models available. The total number of co-crystal sites corresponds to the number of sites in PDBs binding to small organic ligands. Co-crystal sites with less than 15 residues were excluded to avoid counting the numerous non-specific ligand interactions to small surface patches that exist in many PDB structures. The total number of predicted sites on models and PDBs corresponds to sites predicted with $>$ 60% confidence by the SiteSeeker site annotation algorithm incorporated into the TIP. It should be noted that redundancy of co-crystal sites and predicted sites within a given structure was not removed, so in some cases predicted site annotations on PDBs will overlap with co-crystal sites.

Family	# All structures (PDBs+models)	# All PDB Co-crystal sites	# All predicted sites (from PDBs+models)	Average # Predicted sites per structure
ATPases	156	138	426	2.7
Metalloproteases	285	353	538	1.9
Misc. Hydrolases	877	879	1835	2.1
Phosphatases	199	76	301	1.5
Phosphodiesterases	88	126	144	1.6
Trypsin-like Proteases	876	645	1026	1.2
Cyclophilin-type isomerases	124	11	151	1.2
Misc. Other Enzymes	409	353	778	1.9
Misc. Oxidoreductases	657	1406	1414	2.2
P450s	114	27	418	3.7
Short chain dehydrogenases	132	231	245	1.9
Misc. Transferases	622	1279	1438	2.3
Protein Kinases	1010	431	2029	2.0
GPCRs, Family 1	745	80	2595	3.5
GPCRs, Family 2	5	0	2	0.4
Other Cell Surf. Receptors	97	18	157	1.6
Ion Channels	154	123	159	1.0
Ion Pumps	85	9	115	1.4
Mitochondrial transporters	54	14	312	5.8
Other Membrane Proteins	4	0	5	1.3
Nuclear Hormone Receptors	222	325	412	1.9
Other drug target domains	1248	248	1368	1.1
Total	8163	15868	6772	1.9

From druggable proteins to druggable sites

The vast majority of small molecule drugs and drug-like compounds exert their effects by directly interacting with and modulating the activity of a protein target via a specific binding event within a buried pocket or surface cleft of the protein. In light of this observation, another way to characterize the true number of protein targets a small molecule can bind to is by determining the total number of druggable *binding sites* within all members of druggable domain families, since any single target may have multiple sites where small molecule modulators may bind (e.g. substrate sites, cofactor sites, allosteric sites, activator sites, dimer interfaces, or protein-protein interaction sites).

To characterize the number of distinct small molecule binding sites found within the druggable targets for which structures are available, we have analyzed the SiteSeeker ligand-binding site annotations from the TIP knowledgebase. This binding site annotation includes sites which have been experimentally verified to bind small molecules via cocrystallization in PDB structures, as well as sites predicted to bind small molecules based on either a direct mapping from co-crystal sites in PDB structures, or based on geometric and physicochemical properties conducive to binding small molecules. By taking into account all of the predicted and known ligand-binding sites annotated for all druggable target comparative models and PDB structures, we find that there is on average approximately two predicted ligand-binding sites per druggable target structure (Table 2). The observation that multiple ligand binding sites can be found on drug targets is particularly relevant in families where selectivity is of paramount importance, such as the protein kinase family. Since the ATP sites of protein kinases are highly conserved and often extremely difficult to selectively target with small molecule inhibitors, a key strategy that has emerged for kinase drug discovery is the identification of alternate allosteric sites which can be targeted with a much greater degree of selectivity [25].

Site-based identification of "Off-target" opportunities and liabilities

In addition to opening up opportunities for identifying alternative binding sites to target for small molecule inhibition, a "site-centric" view of druggable target space offers key advantages for predicting selectivity and/or cross-reactivity among off-targets. In contrast to the more coarse metrics of sequence and structural similarity, target family-wide com-

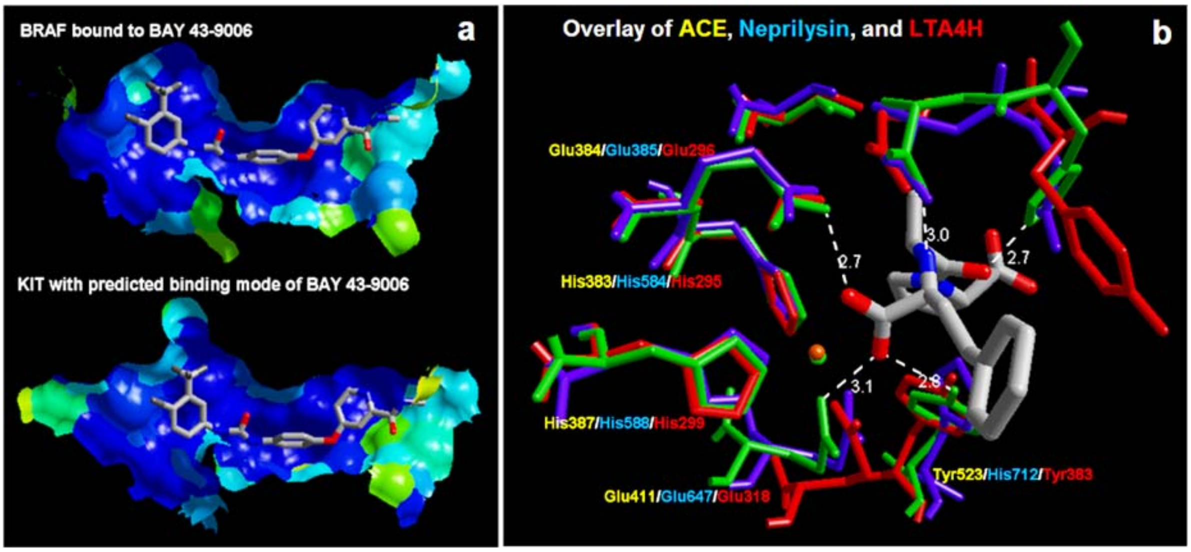

Figure 4. Examples of "off-target" site cross-reactivity between members of the same structural family (a) as well as between members of different structural families (b). (a) The molecule BAY 43-9006 potently inhibits both BRAF and KIT kinase, as well as several other kinases. Shown here are the binding sites of BAY 43–9006 in BRAF and KIT, from both its co-crystal structure with BRAF (PDB 1 uwh) and its predicted binding mode in KIT (PDB 1t46). The proteins' ligand binding site surfaces are colored according to their physicochemical similarity to each other, on a gradient from dark blue (spatially similar positioning of identical functional groups) to yellow (dissimilarity of functional groups). The strong similarity of the binding pockets suggests a high probability of cross-reactivity, even though these proteins share only 30% overall sequence similarity. (b) Angiotensin Converting Enzyme (ACE), Neprilysin, and Leukotriene A4 Hydrolase (LTA4 H) have no overall similarity at the sequence or structural level, however they are all members of the Metalloprotease class of targets, and all have been shown to potently bind ACE inhibitors such as captopril and enalaprilat [26, 27]. Shown here is an overlay of the three active sites, from the ACE-enalaprilat co-crystal structure PDB 1uze, the LTA4H PDB structure 1gw6, and a STRUCTFAST homology model of Neprilysin, highlighting the conserved residues in the zinc-chelating region of the pocket as well as in several regions of the pocket making important contacts with the enalaprilat ligand.

parison of ligand-binding sites allows a much more detailed view of the similarities and differences that are truly relevant for selective drug design. Figure 4 provides two examples where a site-centric view of target similarity offers insight that could not be gained by looking at sequence or structural similarity alone.

The concept of binding site based prioritization of likely off-target liabilities (or conversely, off-target opportunities, if the simultaneous inhibition of multiple targets is a desired effect) is particularly important for large target families such as protein kinases and proteases. For these families, the selection of the *right* off-targets to screen against early on is a challenging task, but potentially a very valuable one given the significant downstream costs likely to be incurred from negative off-target effects. Furthermore, binding site similarity detection methodologies can also extend our understanding of off-target similarities *between* target families that are otherwise completely dissimilar at the structural or sequence level, as shown in the examples in Figures 4b and 5.

Implications for drug discovery

While the overall amount of 3D structural information that can be gleaned for so-called "druggable" human target families is quite impressive, there is still significant room for improvement, particularly in the resolution at which these structures can be annotated. Hence, our analysis underscores the continuing need for additional experimentally derived protein structures for therapeutically relevant target families, an observation in agreement with another recent review [32]. As of November 2005, only about 12.5% of the 5,930 druggable sequence domains described in this work currently have at least one experimental structure available in the PDB. However it should be noted that many targets actually have multiple PDB structures available, often with subtly different 3D conformations or complexed with different ligands (for example CDK2, p38 kinase, Trypsin, COX-2, PDE5). Of the remaining 5,190 targets, roughly 77% can be modeled using comparative modeling approaches, although nearly half of these structures can only be predicted with highly sensitive profile-based homology modeling methods which are capable of extending well into the remote homology "twilight zone" of 15–30% sequence similarity to known PDB structures. It is important to note that a large proportion (roughly 61%) of these twilight-zone targets, as well as targets for which there is no structure available at all, fall into protein families that are membrane-bound. Most notable among this group are the rhodopsin-like GPCR family, the largest and most prolific family of drug targets but a family for which only one suitable crystal structure template is available, bovine rhodopsin [33]. Membrane proteins such as GPCRs and Ion Channels remain the "holy grail" of structural biology, since these proteins are notoriously difficult to crystallize,

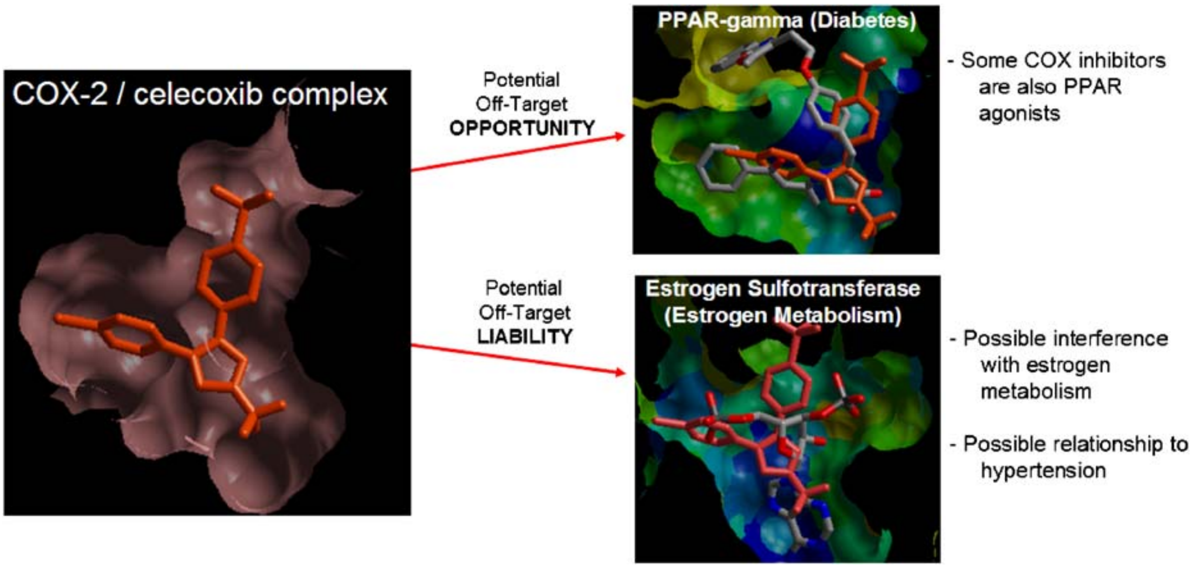

Figure 5. Example of potential opportunities and liabilities that can be exposed through off-target binding site similarity relationships. We performed a site-based similarity search using the COX-2 binding site for the inhibitor celecoxib as the query to retrieve similar "off-target" sites from the full database of druggable binding sites in TIP. Two of the top-ranking sites were the agonist binding site of PPAR-gamma, and the PAP co-factor-binding site of estrogen sulfotransferase (SULT1E1). The top right figure shows celecoxib overlaid into the PPAR-gamma agonist binding-pocket, derived from an optimal overlay of the two sites using the site overlay feature of our EVE software. The binding pocket shapes are well conserved, as are several key binding residues. Interestingly, it has been established in the literature that numerous COX-2 inhibitors do in fact also have PPAR agonist activity [28]. A detailed understanding of the similarity between these two sites offers the off-target *opportunity* that COX inhibitor scaffolds may represent useful starting points for the design of novel PPAR ligands. On the other hand, the similarity to the PAP co-factor binding site of SULT1E1 (bottom right) represents a potential off-target *liability* due to possible interference with normal estrogen metabolism, since SULT1E1 is responsible for the conversion of estradiol to its inactive sulfoconjugate [29]. Furthermore, it is possible that the well publicized hypertensive side effects of COX-2 inhibitors such as Celebrex and Vioxx may be indirectly related to sulfotransferase inhibition. In fact the COX-2 inhibitor etoricoxib (Arcoxia, approved in EU) is known to inhibit SULT1E1 [30], leading to increased serum concentrations of ethinylestradiol. Interestingly, separate studies have been shown that high serum concentrations of ethinylestradiol can contribute to increased fluid retention and hypertension via activation of the renin-angiotensin-aldosterone system [31].

but new structural information for these well-validated drug targets will undoubtedly have the greatest impact on drug discovery.

As the output of high throughput structural genomics initiatives and techniques for membrane protein crystallization continue to improve over the next five to ten years, we can expect the structural coverage of many important, therapeutically relevant gene families to become much more complete. As these new crystal structures are deposited to the public domain to fill in the holes in structural space, the ability to produce more and more accurate comparative models to fill in the remaining gaps will concurrently improve.

With over 30,000 experimental protein structures currently in the PDB, and approximately 100 new crystal structures being deposited to the PDB *per week*, it is becoming increasingly important to develop efficient infrastructures for properly storing, annotating, mining, and analyzing this deluge of structural data, in much the same way that analogous infrastructures have been developed to handle large amounts of sequence, chemical structure, and gene expression data in the disciplines of bioinformatics, chem-informatics, and proteomics, respectively. As the amount of structural data increases, structural informatics databases and data-mining tools such as the Target Informatics Platform will be necessary to extract the maximum value from every new experimental structure that is solved. Most importantly, structural informatics approaches for analyzing ligand binding site similarities on a proteome-wide scale have particular value in identifying off-target cross-reactivity, giving discovery researchers powerful insight into potential risks and opportunities as early in discovery as possible.

References

1. Venter, J.C., et al., *The sequence of the human genome*, Science, 291 (2001) 1304–1351.
2. Hopkins, A.L. and Groom, C.R., *The druggable genome*, Nat. Rev. Drug. Discov., 1 (2002) 727–730.
3. Zambrowicz, B.P. and Sands, A.T, *Knockouts model the 100 best-selling drugs–will they model the next 100*? Nat. Rev. Drug. Discov., 2 (2003) 38–51.
4. http://www.structuralgenomics.org/
5. http://www.sgc.utoronto.ca/
6. Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C, Murzin, A.G., *SCOP database in 2004: Refinements integrate structure and sequence family data*, Nucleic Acids Res., 32 (2004) D226–D229.
7. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M., *CATH – a hierarchic classification of protein domain structures*, Structure, 5 (1997) 1093–1108.
8. Holm, L., Sander, C., *The FSSP database of structurally aligned protein fold families*, Nucleic Acids Res., 22 (1994) 3600–3609.

9. Buchan, D.W., Rison, S.C., Bray, J.E., Lee, D., Pearl, F., Thornton, J.M., Orengo, C.A., *Gene3D: Structural assignments for the biologist and bioinformaticist alike*, Nucleic Acids Res., 31 (2003) 469–473.
10. Gibrat, J.F., Madej, T., Bryant, S.H., *Surprising similarities in structure comparison*, Curr. Opin. Struct. Biol., 6 (1996) 377–385.
11. Kersey, P.J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., Apweiler, R., *The international protein index: An integrated database for proteomics experiments*, Proteomics, 4 (2004) 1985–1988.
12. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., *The protein data bank*, Nucleic Acids Res., 28 (2000) 235–242.
13. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., *Basic local alignment search tool*, J. Mol. Biol., 215 (1990) 403–410.
14. Debe, D.A., Danzer, J.F., Goddard, W.A. 3rd and Poleksic, A., *STRUCTFAST: Extreme remote homology detection and alignment using novel dynamic programming and profile-profile scoring*, Proteins, submitted.
15. Xie, L., Danzer, J.F. and Debe, D.A., *publication in progress*.
16. Xie, L., Danzer, J.F. and Debe, D.A., *publication in progress*.
17. Schmitt, S., Kuhn, D., Klebe, G., *A new method to detect related function among proteins independent of sequence and fold homology*, J. Mol. Biol., 323 (2002) 387–406.
18. Palmer, B., Danzer, J.F., Hambly, K. and Debe, D.A., *StructSorter: A continuously maintained pair-wise structure alignment of a comprehensive protein structure database*, Bioinformatics, submitted.
19. Mulder, N.J., et al., *InterPro: An integrated documentation resource for protein families, domains and functional sites*, Brief. Bioinform., 3 (2002) 225–235.
20. Lipinski, C.A., *Drug-like properties and the causes of poor solubility and poor permeability*, J. Pharmacol. Toxicol. Methods, 44 (2000) 235–249.
21. Bateman, A., et al., *The Pfam protein families database*, Nucleic Acids Res, 32 (2004) D138–D141.
22. McGinnis, S. and Madden, T.L., *BLAST: At the core of a powerful and diverse set of sequence analysis tools*, Nucleic Acids Res., 32 (2004) W20–W25.
23. Hillisch, A., Pineda, L.F. and Hilgenfeld, R., *Utility of homology models in the drug discovery process*, Drug Discov. Today, 9 (2004) 659–669.
24. Dahl, S.G. and Sylte, I., *Molecular modelling of drug targets: The past, the present and the future*, Basic Clin. Pharmacol. Toxicol., 96 (2005) 151–155.
25. Parang, K. and Sun, G., *Design strategies for protein kinase inhibitors*, Curr. Opin. Drug Discov. Devel., 7 (2004) 617–629.
26. Dumoulin, M.J., Adam, A., Rouleau, J.L. and Lamontagne, D., *Comparison of a vasopeptidase inhibitor with neutral endopeptidase and angiotensin-converting enzyme inhibitors on bradykinin metabolism in the rat coronary bed*, J. Cardiovasc. Pharmacol., 37 (2001) 359–366.
27. Thunnissen, M.M., Andersson, B., Samuelsson, B., Wong, C.H. and Haeggstrom, J.Z., *Crystal structures of leukotriene A4 hydrolase in complex with captopril and two competitive tight-binding inhibitors*, Faseb. J., 16 (2002) 1648–1650.
28. Bernardo, A., Ajmone-Cat, M.A., Gasparini, L., Ongini, E. and Minghetti, L., *Nuclear receptor peroxisome proliferator-activated receptor-gamma is activated in rat microglial cells by the anti-inflammatory drug HCT1026, a derivative of flurbiprofen*, J. Neurochem., 92 (2005) 895–903.
29. Tanaka, K., Kubushiro, K., Iwamori, Y., Okairi, Y., Kiguchi, K., Ishiwata, I., Tsukazaki, K., Nozawa, S., Iwamori, M., *Estrogen sulfotransferase and sulfatase: Roles in the regulation of estrogen activity in human uterine endometrial carcinomas*, Cancer Sci., 94 (2003) 871–876.
30. European Medicines Agency; Summary of Product Characteristics – Etoricoxib. http://www.emea.eu.int/pdfs/human/epar/Etoricoxib.pdf.
31. Elger, W., *Conception and pharmacodynamic profile of drospirenone*, Steroids, 68 (2003) 891–905.
32. Mestres, J., *Representativity of target families in the Protein Data Bank: Impact for family directed structure-based drug discovery*, Drug Discov. Today, 10 (2005) 1629–1637.
33. Palczewski, K., et al., *Crystal structure of rhodopsin: A G protein-coupled receptor*, Science, 289 (2000) 739–745.