Chemical Fragments as Foundations for Understanding Target Space and Activity Prediction

Jeffrey J. Sutherland,[†] Richard E. Higgs,[‡] Ian Watson,[§] and Michal Vieth*[,§]

Discovery Informatics, Discovery Statistics, and Discovery Chemistry of Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, Indiana 46285

Received November 7, 2007

The use of small inhibitors' fragment frequencies for understanding kinase potency and selectivity is described. By quantification of differences in the frequency of occurrence of fragments, similarities between small molecules and their targets can be determined. Naive Bayes models employing fragments provide highly interpretable and reliable means for predicting potency in individual kinases, as demonstrated in retrospective tests and prospective selections that were subsequently screened. Statistical corrections for prospective validation allowed us to accurately estimate success rates in the prospective experiment. Selectivity relationships between kinase targets are substantially explained by differences in the fragment composition of actives. By application of fragment similarities to the broader proteome, it is shown that targets related by sequence exhibit similar fragment preferences in small molecules. Of greater interest, certain targets unrelated by sequence are shown to have similar fragment preferences, even when the chemical similarity of ligands active at each target is low.

Introduction

The recognition of fragments in small organic molecules is an intuitive process for medicinal chemists, reflecting the manner in which molecules are synthesized from chemical building blocks. Algorithmic approaches for reducing molecules to fragments have been used for retrosynthetic analysis[1,2] and for the characterization of their druglike properties.[3,4] In addition, protein receptors recognize certain fragments with greater preference. The concept of "privileged structure" is grounded in the observation that certain substructures confer potency within a class of targets;[5–8] hydroxamates confer potency for matrix metalloproteases, benzamidine for serine proteases, and aminopyrimidines for kinases and ATP binding proteins. The linking of privileged fragments with fragments from other sources (e.g., drugs) is a well-established approach for designing targeted chemical libraries[9] (Figure 1). Understanding target similarity relationships via fragments has strong appeal in fragment screening and hit generation via fragment growing and linking.[10,11] Such fragment centric relationships offer the possibility of tailoring screening sets for a given target using greater wealth of data at a related target.

The properties of small molecules have been used to create pharmacological maps, which describe relationships among proteins. Our earlier work defined SAR similarity, which relates the extent to which compounds assayed at two targets inhibit them in a similar manner.[12,13] The relationship between targets and their ligands has been described by Frye.[14] Paolini and colleagues employ simpler measures of cross-target promiscuity,[15] which also require affinity determination of compounds at both targets. Izrailev and Farnum[16] have demonstrated the potential to annotate targets of unknown function by assessing the similarity of ligands for the query target to those of well-annotated reference targets. Bender and colleagues used principal component analysis on multitarget affinity predictions for defining bioactivity spaces.[17,18] An approach described by Keiser and colleagues[19] circumvents the need for activity data at two targets; Daylight fingerprints coupled to a statistical model can be shown to group targets via their ligands in a biologically meaningful way.

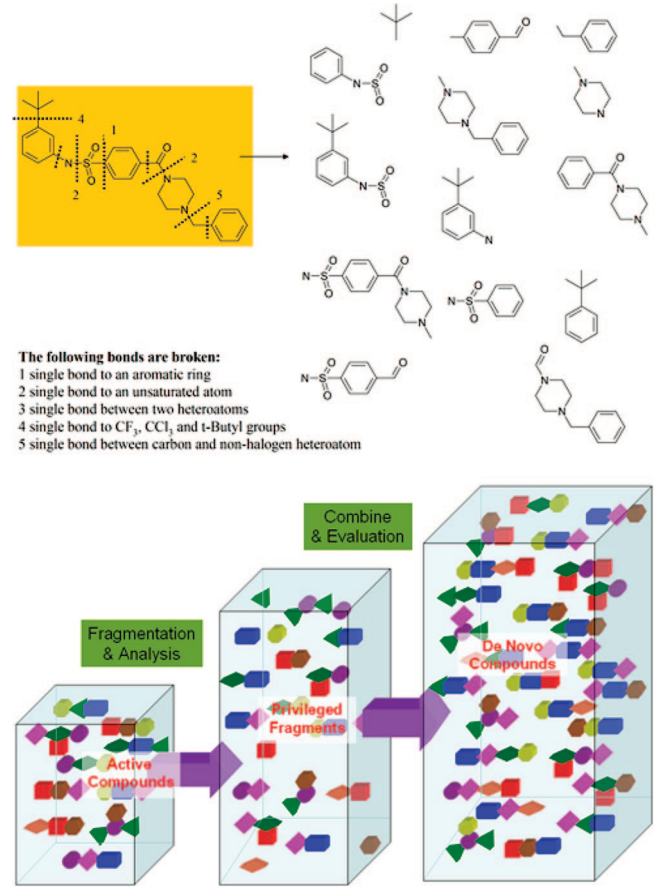


Figure 1. (Top) Schematic description of the Dicer algorithm for reducing small organic molecules to fragments. (Bottom) The reduction of molecules to fragments, coupled to chemically intelligent reassembly rules, allows for the enumeration of large virtual chemistry spaces.

* To whom correspondence should be addressed. Phone: 317-277-3959. Fax: 317-276-6545. E-mail: m.vieth@lilly.com.
[†] Discovery Informatics.
[‡] Discovery Statistics.
[§] Discovery Chemistry.

10.1021/jm701399f CCC: $40.75
Published on Web 04/04/2008

In this work, we describe a simple approach for reducing organic molecules to a series of redundant fragments (Figure 1), using five simple rules motivated by the manner in which chemists reduce molecules to reagents. Encoding the properties of small molecules via their fragment composition allows for reliable affinity prediction using standard cheminformatics methods while retaining the intuitive characteristics of fragments. We show that the frequency at which fragments occur in active molecules can be used to define target similarity, in a manner that complements sequence-based comparisons and the whole-molecule approaches listed above. The relationship between fragment frequencies in active compounds and kinase selectivity is demonstrated. Fragment similarities can be applied to targets spanning the proteome,[20] identifying targets unrelated in sequence that show preference for similar chemical fragments, even in cases where the overall similarity of active compounds is low.

Predicting Kinase Activity from Fragment Composition

The use of fingerprint representations of small molecules is widespread in chemical informatics. Daylight fingerprints (www.daylight.com) and related methods enumerate paths of various length through molecules, using the types and number of such paths to set various bits in a fingerprint. MACCS fingerprints (www.mdl.com) use a simpler approach, noting the presence or absence of functional groups within a predefined dictionary. An implementation of Merck atom-pair fingerprints[21] enumerates pairs of atoms in molecules at various bond separations; the atoms are distinguished using topological torsion descriptors[22] (MAPTT[a]). There are other fingerprinting approaches,[23] e.g., 3D pharmacophore fingerprints, that we have not examined here. A common aspect of fingerprint approaches is the difficulty of casting the bit-string representation into chemically intelligible form. For the examples above, it is most severe for Daylight fingerprints (i.e. it is hard to know what a "1" bit means).

The reduction of small organic molecules to fragments with the Dicer algorithm (Methods) allows for their representation via fragment fingerprints. A fragment-derived fingerprint denotes the presence or absence of each fragment within a particular molecule. We define fingerprints of moderate length by considering only fragments occurring more than 1 time per 1000 molecules within the set of compounds under consideration. For the kinase set discussed below, this results in approximately 1800 fragments.

Beyond their traditional use in similarity searching, chemical fingerprint methods can be used for QSAR modeling and activity prediction (e.g., using partial least-squares[24] or naive Bayes models[15,17,18]) in which the weight of each chemical feature is adjusted with reference to the known activities of training compounds. Because of their simplicity and ease of interpretation, we explore the usefulness of naive Bayes models, in which compounds are described by Daylight, MACCS, MAPTT, and fragment fingerprints. For each fingerprint bit or fragment, its presence in a given molecule is denoted by 1 (present) or 0 (absent).[25] A naive Bayes model indicates the probability of observing activity, given the presence or absence of chemical features in molecules (Methods).

Thirty-six kinase assays were assembled from our internal research efforts, with each kinase having at least 500 actives ($\geq$70% inhibition at 20 μM test substance concentration)[26] and 2000 or more compounds assayed. Models were constructed using 67% of actives and inactives, with the remainder used for model validation; compounds were randomly allocated to training and test sets, and results were averaged over 10 such splits. The usefulness of models was quantified using two approaches. A conventional measure of success in virtual screening is the enrichment ratio (ER); it indicates the relative improvement in the number of active molecules within the top *X*% of a set ranked by score, compared to random selection. We report enrichment ratios at the top 1, 5 and 10% of compounds ranked by score (Table 1).[27] On average, fragment-based models exceed the predictive accuracy of Daylight, MACCS, or MAPTT models in retrospective cross-validation.

Table 1. Comparing Fingerprints for Single Point Activity Prediction: Average (Standard Deviation) over 36 Kinases and 10 Train/Test Splits Per Kinase

	Daylight	MACCS	MAPTT	fragments	maximum possible[a]
ER1%	5.1 (0.81)	5.9 (0.87)	5.8 (0.85)	8.9 (0.58)	11.1
ER5%	5.1 (0.32)	4.0 (0.42)	5.2 (0.35)	6.7 (0.28)	11.0
ER10%	4.6 (0.20)	3.0 (0.26)	4.2 (0.21)	5.1 (0.17)	6.7
AUC	0.81 (0.01)	0.70 (0.02)	0.79 (0.01)	0.87 (0.01)	1.0

[a] Maximum enrichment possible, obtained by selecting only actives until there are none left.

An alternative to enrichment ratio calculations consists of plotting the true positive rate (sensitivity) vs 1-true negative rate (1-specificity) while varying the threshold used for classifying compounds as active or inactive; i.e., a "receiver-operator characteristic" (ROC) plot. For such ROC plots, the area under the curve (AUC) represents the probability that a randomly selected active compound scores higher than a randomly selected inactive. The ranking of methods mirrors that obtained with conventional enrichment ratios. Models using fragments outperform those using Daylight, MACCS, and MAPTT fingerprints in the ER and AUC metrics. The overall ranking of fragments > Daylight > MAPTT > MACCS is observed for most kinases except for the insulin receptor kinase (INSR), where the fragment model has a lower AUC value than the Daylight model (AUC 0.79 vs 0.81) (Figure 2).

The fragment naive Bayes models for six kinases were used in a prospective virtual screening of our compound collection, with the goal of identifying new/previously undiscovered actives. Compounds having a probability score greater than 0.5 in any model were selected and tested at all six kinases, using the same assays employed for training and validating the models.

It is occasionally remarked that predictive models fail to meet expectations; the source of those expectations are often in the assumptions made in retrospective cross-validation.[28] One factor responsible for discrepancies between expected and observed accuracy is the different manner in which the compounds used for developing the model are selected for screening. In this work, most assayed compounds were synthesized as libraries designed to inhibit various kinases, were purchased with a similar intent, or were selected from lead optimization efforts against kinase targets. The prevalence of actives in such a set is much higher than prevalence of actives in our collection as a whole. As such, the sensitivity (true positive rate) and specificity (true negative rate) determined through cross-validation on previously assayed compounds convey excessively optimistic expectations of hit rates obtained in virtual screening applications on general compound collections, since the composition of the compound sets differs markedly.

The positive predictive value (PPV) of a model, otherwise referred to as model hit rate or precision, is the probability that

[a] Abbreviations: MAPTT, Merck atom pair/topological torsion fingerprints; ER, enrichment ratio; TP, true positive rate; TN, true negative rate; ROC, receiver-operator characteristic; AUC, area under the curve; PPV, positive predictive value; GPCR, G-protein-coupled receptor; NHR, nuclear hormone receptor; DIOD, drugs in other drugs; RECAP, retrosynthetic combinatorial analysis procedure.

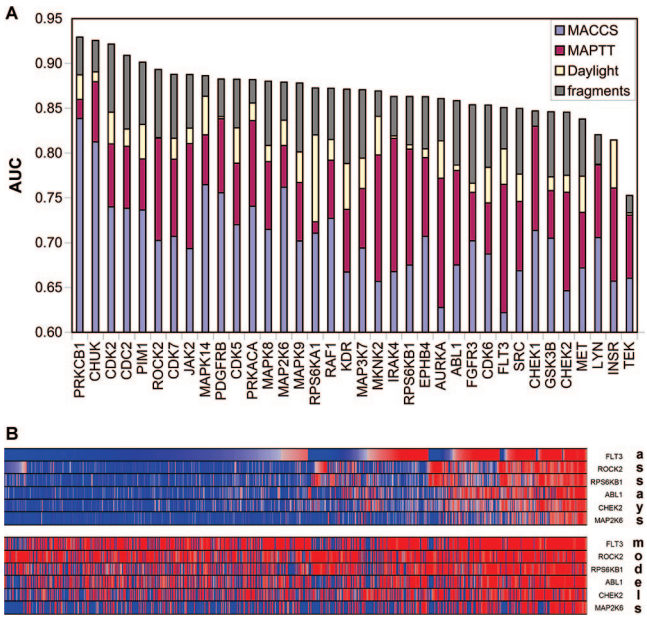


Figure 2. Predictive models of kinase activity. (A) Area under the curve (AUC) from receiver-operator characteristic (ROC) plots determined with retrospective validation of naive Bayes models using fragments, Daylight, MACCS, and MAPTT fingerprints. Larger AUC values (maximum 1) indicate superior models. For 35/36 kinases (i.e., INSR excepted), fragment models have larger AUC values than the other methods. Two-sided *t* tests indicate that all differences between fragment models and those using other descriptors are significant ($P < 0.05$). (B) Comparison of scores from fragment naive Bayes models (range 0−1) to assay results (range 0−100% inhibition at 20uM) for six kinases screened in prospective tests. Red indicates activity with blue representing lack thereof.

a compound predicted to be active is found to be active. It can be calculated from the sensitivity, specificity, and estimated prevalence (Methods). While this measure is infrequently used in virtual screening application, it is a standard measure used in many disciplines for conveying the "real world" expectations of a diagnostic test or model. At low prevalence (i.e., hit rate) of actives across the entire screening collection, the specificity of models affects the PPV significantly more than sensitivity, since models will be evaluating inactive compounds far more frequently than actives. Typical hit rates from medium- and high-throughput screens (in which the selection of compounds from the collection is significantly more random) are 0.1–10%. The hit rate observed in screening may be lower than the PPV, since the latter does not account for factors such as extrapolation from training data, errors in structures, etc.

Comparisons between retrospective cross-validation and results from prospective screening against the six kinases are given in Table 2 and Figure 2. The large difference in specificity between retrospective and prospective studies can be understood from the manner in which compounds were selected: the specificity for the prospective experiment could have been improved by screening more of the compounds predicted inactive, since most compounds in our collection will be inactive at a given target. Correspondingly, the sensitivity would decrease, since some of the untested compounds will be found active, despite being predicted inactive. Of 1965 compounds predicted active and tested, 42% tested active at one or more kinases; 87% of 30 compounds predicted inactive at every kinase test inactive at every kinase. Our model hit rates significantly exceed the average prevalence of actives in the training data (24%, i.e., synthesizing/purchasing compounds without using a predictive model) and typical hit rates from screening campaigns.

PPVs calculated using the prevalence of actives in the training and test data are approximately twice those obtained in the prospective screening experiment (PPV for the latter is simply the fraction of predicted actives that confirmed active in the assay). Because of differences between the training/test compounds and our corporate collection, we calculate PPVs using two estimates of prevalences in our collection: (1) assuming that each kinase has a 5% prevalence of actives and (2) using the actives prevalence observed when screening diverse compound cassettes at that target. The exact value of prevalence can only be determined by screening the entire set of compounds in the collection (which would defeat the purpose of predictive modeling). Calculating PPV using a few plausible prevalence values observed from similar screens serves to highlight the success rate that can be expected from prospective screening.

Understanding Kinase Targets with Fragments

A number of reports have described the concept of structure–activity relationship (SAR) similarity of kinases, a metric that quantifies the degree to which compounds assayed at two kinases exhibit similar activity in each assay (i.e., cross-activity) and reflects the ease or difficulty of obtaining selectivity at one target vs another. In this work, we calculate the SAR similarity of two kinases using screening data obtained from internal profiling of compounds at 53 kinases (the previous 36 kinases and an additional 17 having fewer than 500 actives per target). Each assay has been used to profile 2000 or more compounds, and each pair of assays has 570 or more compounds in common. The SAR similarity of two kinases is strongly (negatively) correlated with the average activity difference calculated over all compounds assayed at both targets (Pearson correlation coefficient r for SAR similarity vs average activity rmsd = −0.89 for 1378 pairs of 53 kinases).

We explore the use of fragment frequencies as predictors for SAR similarity by quantifying similarities in the distribution of fragments within actives at each kinase (i.e., with no requirement that the actives at one target have been assayed at the other; actives have ≥70% inhibition at 20 μM screening concentration). Fragment similarity between two kinases is calculated with the Tanimoto coefficient (T_c), using the frequency of approximately 1800 fragments observed in the tested compounds (T_c ranges between 0 and 1, with 1 indicating identical fragment frequencies observed for both proteins).

The SAR similarity of kinases is reasonably predicted by the fragment composition of actives. To compare fragment composition with more established chemical fingerprinting methods, target similarities were calculated using Daylight fingerprints and MACCS keys using the same procedure (i.e., calculating the frequency of each Daylight bit or MACCS key in actives for a given target and comparing two targets via the Tanimoto coefficient). Fragment frequencies are substantially more predictive of the experimentally determined SAR similarity of kinases than either Daylight fingerprints or MACCS keys (Figure 3). In a small number of cases, kinases having high sequence similarity are found to have fragment similarities outside the highest quartile: EphB4 vs EphA4 (81%), PKCe vs PKCh (78%), and PKCh vs PKCb2 (68%) have high sequence similarity (indicated in parentheses) but lower fragment similarity; all have high SAR similarity, indicating shortcomings in the fragment similarity metric.

The determination of SAR similarity, or any other measure of target affinity correlation, requires a substantial quantity of

Table 2. Retrospective Validation and Prospective Screening Results Using Fragment-Based Naive Bayes Models

	FLT3	ABL1	ROCK2	RPS6KB1	CHEK2	MAP2K6	any
Retrospective Cross-Validation							
training data prevalence[a]	0.24	0.11	0.12	0.13	0.10	0.07	0.24
sensitivity	0.59	0.56	0.59	0.57	0.51	0.57	
specificity	0.89	0.93	0.92	0.90	0.91	0.95	
PPV@test set prev[b]	0.64	0.51	0.52	0.47	0.39	0.45	
PPV@5%[c]	0.22	0.31	0.30	0.24	0.23	0.36	
diverse cassette prevalence[d]	0.15	0.03	0.05	0.08	0.04	0.03	
PPV@diverse cassette prevalance[e]	0.49	0.22	0.26	0.34	0.19	0.25	
Prospective Screening Results (1995 Compounds Tested)							
no. active (% active)	478 (24.0)	320 (16.0)	318 (15.9)	302 (15.1)	251 (12.6)	150 (7.5)	828 (41.5)
no. pred active	1468	1190	1574	1260	1079	664	1965
sensitivity	0.93	0.81	0.96	0.90	0.85	0.87	1.00
specificity	0.33	0.44	0.24	0.42	0.50	0.71	0.02
PPV[f]	0.30	0.22	0.19	0.22	0.20	0.20	0.42

[a] The prevalence of actives in the data used for training and cross-validating models. [b] PPV calculated using training/test set prevalence. [c] PPV assuming a 5% prevalence of actives in our collection. [d] Actives prevalence observed when screening diverse compound cassettes. [e] PPV calculated using actives prevalence from diverse compound cassettes. See Discussion for why PPV using cross-validation test set prevalence is a poor predictor of prospective screening PPV. [f] PPV for the prospective screening results is simply the number of compounds predicted active that were confirmed active divided by number of compounds predicted active (i.e., the confirmation rate in the assay).

affinity measurements from compounds assayed at both targets. For this reason, a number of approaches have examined the possibility of understanding SAR similarity in the context of more directly accessible quantities, such as sequence similarity[12] and active-site similarity.[13,31,32] The fragment similarity defined in this work requires only a modest number of compounds (~50) active at each target, with no requirement for testing the same compounds at both targets. The determination of SAR similarity itself requires a number of assumptions to be made, including the minimum number of compounds assayed at both targets, what constitutes a reasonably diverse compound set, and the type of similarity metric (e.g., Tanimoto, cosine, Euclidian distance, etc.).

In Figure 3, we compare SAR similarity determined using our entire tested compound set to an analogous quantity from the subset appearing in PubChem (pubchem.ncbi.nlm.nih.gov), our previous work[13] that employed actives with IC_{50} data from the literature, and published data from Ambit Biosciences[29] on 19 clinical/launched inhibitors. The modest correlations for the last two sources highlight the limitations of using data from the literature where cross-activity data are sparse or of using very small compound sets. The dehydron approach of Fernandez and co-workers[30,31] and our own structure-based method[13] fail to explain our profiling SAR similarity and are substantially worse than a simple measure of sequence identity. In summary, the fragment composition of actives is by far the best predictor of the experimentally determined SAR similarity.

Understanding Activity across Gene Families with Fragments

The utility of fragments for understanding relationships between proteins was examined using a large data set of compounds active across the druggable proteome.[20] The journal and patent literature (as curated by GVK Biosciences, Inc.) contains approximately 151 000 compounds having an IC_{50} or K_i value of ≤ 1 μM at a human, mouse, or rat protein receptor (about 193 000 reported compound-activity values, after retaining only the human value where data exist and using mouse or rat values otherwise. By mapping of the mouse and rat proteins to their human orthologues, 518 proteins have a known ligand with submicromolar activity. In this analysis, we restrict ourselves to the 223 proteins with 100 or more small molecules having activity of ≤ 1 μM.

Compound−target pairs (with targets annotated using official Entrez gene symbols) were assigned to a gene family/subfamily using the Proteome database.[33] The active compounds were reduced to fragments with the Dicer algorithm (Methods). The most frequently occurring small fragments fail to differentiate actives across gene families; phenyl, benzyl, and assorted simple one-member cycles account for approximately 6 of the 10 top fragments (Supporting Information). In contrast, focusing on larger fragments (>8 heavy atoms) identifies recognizable chemical motifs: a GPCR privileged substructure, a fibrate group from PPAR NHR receptors, quinoline and indolyl maleimide groups from kinase inhibitors, and benzamidine from serine proteases (Table 3). Some frequently occurring fragments in active molecules appear not because of their effect on binding affinity but because of their common use for modifying physicochemical properties (e.g., morpholino solubilizing groups used for kinase inhibitors). The unexpected presence of certain fragments arises because of activity across gene families: e.g., biphenyls for ion channel actives arise from GPCR ligands that bind to the KCNH2 and other potassium channels. Comparing the average value of Lipinski properties[34] in fragments and parent molecules across gene families gives a Pearson correlation coefficient (r) of 0.72 for ClogP and 0.89 for the number of hydrogen bond acceptors. Thus, hydrophobic fragments make hydrophobic molecules, as might be expected. There is no correlation for molecular weight, suggesting that building blocks of similar size can be used for creating active molecules across different gene families.

The similarity of proteins can be quantified with a number of methods, including sequence-similarity, protein active site similarity,[13,31,32] affinity differences among a series of molecules assayed at both proteins,[13,15] and the properties of small molecules that bind to the proteins.[16,19] We employ the fragment similarity (described above for understanding kinase selectivity) for establishing the similarity between proteins representing all major gene families from the proteome. The 1200 fragments used for calculating similarities occur more than 1 time per 1000 molecules among the 151 000 active compounds considered.

On average, the fragment similarity of proteins mirrors their sequence similarity; proteins belonging to the same subfamily (e.g., aspartic proteases, class A adrenergic GPCRs, receptor tyrosine kinases) have higher fragment similarities than proteins from the same gene family or different families (Figure 4).

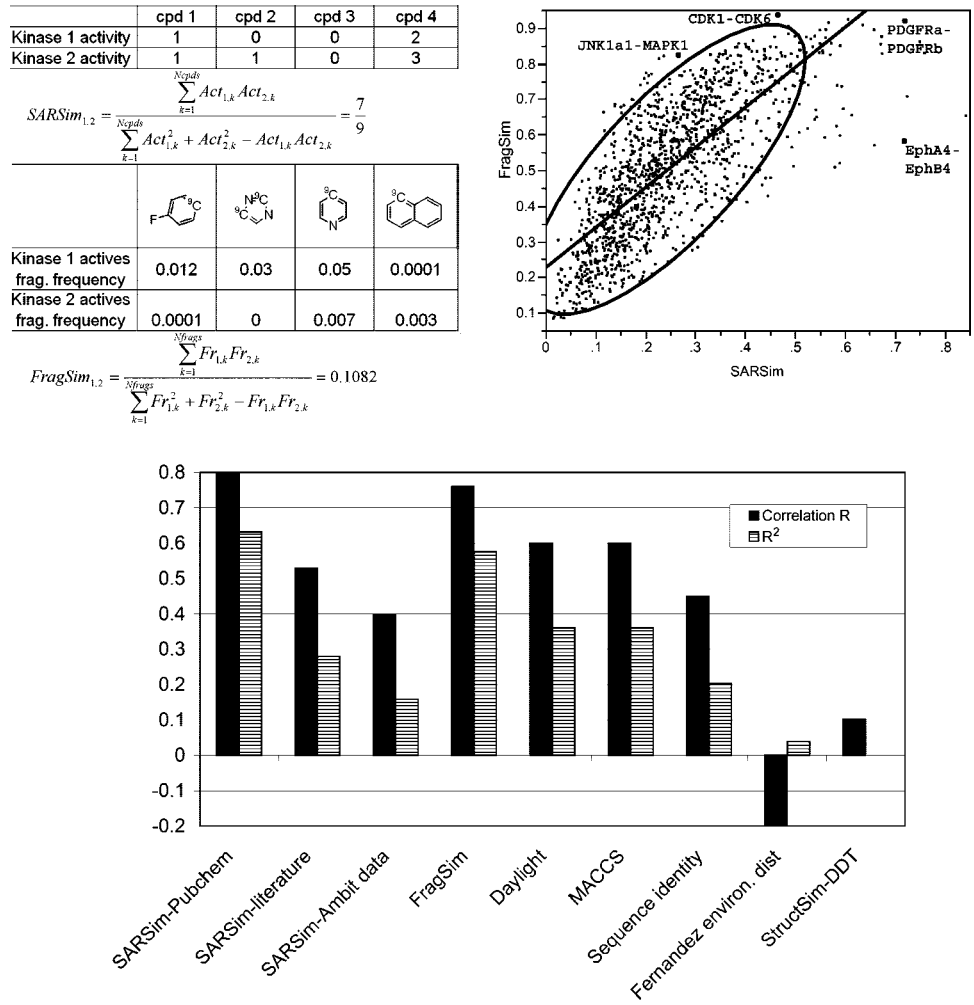
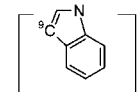
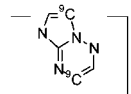
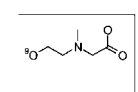
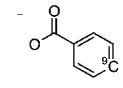
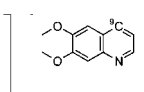
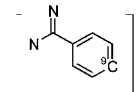
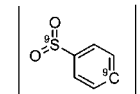
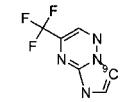
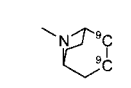
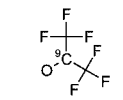
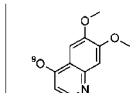
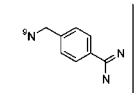
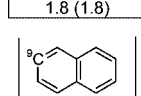
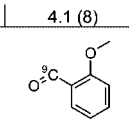
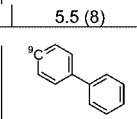
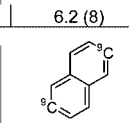
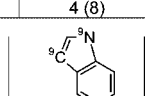
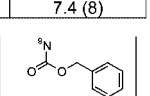
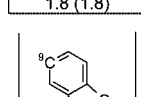
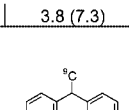
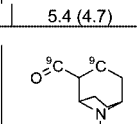
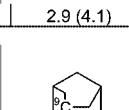
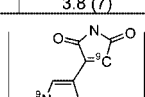
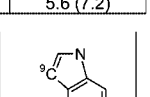
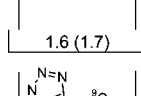
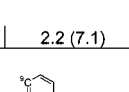
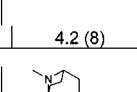
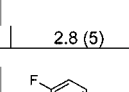
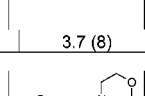
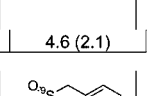
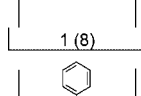
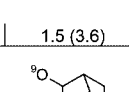
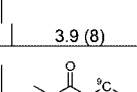
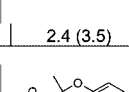
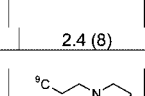
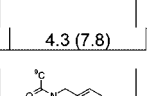
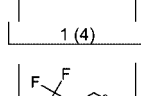
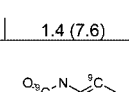
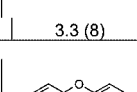
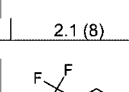


Figure 3. Comparison of SAR (experimental) similarity with various predictors. (Top left) Exemplification of methodology used for determining SAR similarity from compounds assayed at each of two kinases and for calculating kinase similarities from the fragment composition of active compounds (Methods). (Top right) Comparison of SAR similarity from internal profiling vs similarities calculated from fragment frequencies of active molecules. (Bottom) Correlations between SAR similarities from internal profiling and (1) the subset of compounds appearing in PubChem, (2) 291 overlapping kinase pairs from our previous work using IC_{50} data from the literature,[13] (3) using Ambit data published for 19 compounds (staurosporine excluded) at 26 of the 120 reported kinases present on our profiling panel,[29] (4−6) kinase similarities calculated using fragment, Daylight bit, and MACCS key frequencies in actives (Methods), (7) using sequence identity over the catalytic domain, (8) from the dehydron work of Fernandez and co-workers,[30] and (9) using the structure-based approach reported in our previous work.[13] The negative correlation for (8) is expected, since a similarity and a distance are compared. When calculating a SAR distance (Euclidian) rather than a Tanimoto SAR similarity on our profiling data, the correlation (r) vs (8) is 0.32.

Sequence similarity between targets is highest within subfamilies and insignificant between gene families. Hierarchical clustering of fragment-derived protein similarities can reveal target pairs modulated by small molecules containing similar fragments, some of which are not related by sequence (Figures 4). A sample of high-similarity protein pairs from different gene families is shown in Table 4. Certain target pairs, such as the serotonin transporter and serotonin receptor 2A, can be readily rationalized. High fragment similarity of targets does not imply high similarity of ligands, as suggested by the exemplified ligands. Including those shown in Table 4, there are 99 protein pairs from different gene families having fragment similarity greater than 0.5 (i.e., more than 3 standard deviations from the average similarity of proteins from different gene families). Of all compound pairs formed by taking one active at the first target and one active at the second sequence-unrelated target, only 0.004% (1498) have Daylight similarity greater than 0.75 (a lower similarity value than typically used in similarity searching applications). Thus, targets unrelated in sequence having high fragment similarity do not generally bind similar molecules. This can be viewed as having common building blocks in ligands for both targets, which are assembled differently and include fragments that are not shared. Targets having binding pockets with different characteristics can accommodate common fragments.[14] It is important to note that coverage of the proteome by public domain sources is very uneven: there are substantially fewer protease targets having 100 or more actives compared to GPCRs (32 vs 110), which may not accurately represent the true diversity of actives. The general applicability of target maps obtained from ligands should be understood with these data set biases in mind.

The similarity of proteins assessed using fragment frequencies differs from that obtained using the promiscuity index of Paolini and colleagues,[15] which quantifies the ability of molecules to bind at each of two receptors (Figure 5). Significant differences are observed for ligands of class A GPCRs, which are promiscuous within the class and across many others; in contrast, class A GPCRs have lower overall fragment similarity within and across classes. Fragment similarities indicate that protease

Table 3. Most Frequent Fragments with More than Eight Heavy Atoms by Gene Family[a]

GPCR	Ion channel	Transporter	NHR	Kinase	Protease
					
8.2 (3.8)	5.9 (8)	9.7 (8)	12 (4.9)	4.1 (8)	8 (7.7)
					
1.8 (1.8)	4.1 (8)	5.5 (8)	6.2 (8)	4 (8)	7.4 (8)
					
1.8 (1.8)	3.8 (7.3)	5.4 (4.7)	2.9 (4.1)	3.8 (7)	5.6 (7.2)
					
1.6 (1.7)	2.2 (7.1)	4.2 (8)	2.8 (5)	3.7 (8)	4.6 (2.1)
					
1 (8)	1.5 (3.6)	3.9 (8)	2.4 (3.5)	2.4 (8)	4.3 (7.8)
					
1 (4)	1.4 (7.6)	3.3 (8)	2.1 (8)	2.1 (6.4)	4.1 (8)
					
0.9 (1.7)	1.1 (4.8)	2.5 (4.8)	1.6 (2)	2 (7.8)	3.7 (3.8)
0.9 (8)	0.9 (0.8)	2.1 (2.5)	1.2 (3)	1.9 (4.9)	2.9 (7.3)
0.9 (6.7)	0.9 (2.5)	2 (7.1)	1.1 (7.8)	1.7 (7.5)	2.9 (2.9)
0.9 (8)	0.8 (1)	1.9 (4)	0.9 (2.7)	1.5 (8)	2 (7.9)

[a] The numbers below fragments indicate their rate of occurrence per 100 actives from the given gene family, and the value in parentheses is the average rate of occurrence across gene families. Detailed results for the 1000 most frequently-occurring fragments for each gene family are given in the Supporting Information. The total number of active compounds for each gene family is as follows: GPCR 113 770; ion channel 8621; transporter 4479, NHR 6226; kinase 24805; protease 28 895.

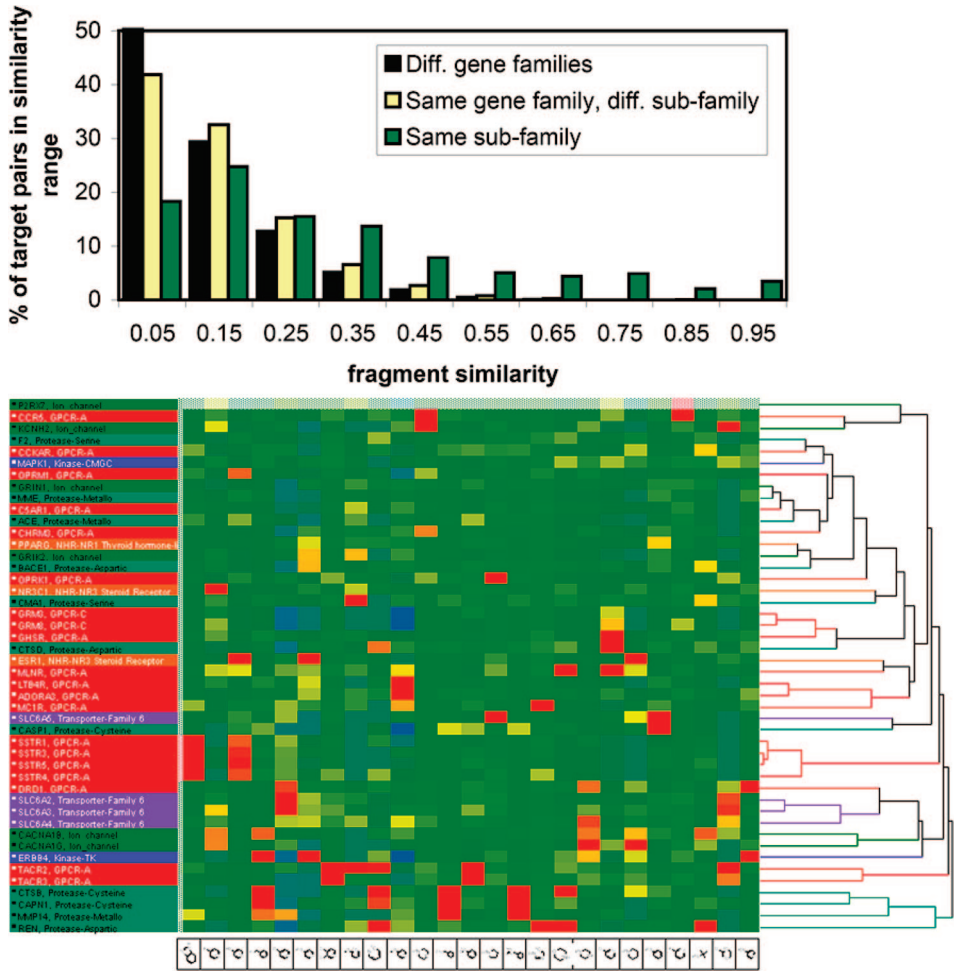


Figure 4. Tanimoto similarities of proteins calculated using fragment frequencies in active molecules ($IC_{50}/K_i \leq 1$ μM). (Top) Distribution of fragment similarities for protein pairs belonging to different gene families (e.g., GPCR vs kinase), the same gene family but different subfamilies (e.g., cysteine vs aspartic proteases), or the same subfamily. The median (75th percentile) for each grouping is 0.10 (0.18), 0.12 (0.20), and 0.24 (0.42) for the three groups. (Bottom) Hierarchical clustering (Ward's method) on a subset of fragment frequencies for 46 targets having fragment similarity to a target of different gene family greater than 0.5; 0.5 was selected on the basis of average similarity + 3 standard deviations for targets from different gene families. Only 25 fragments are shown for clarity; these contain more than four atoms, have frequency of occurrence greater than 15 per 1000 molecules, and contain one or more heteroatoms.

fragments occur frequently in compounds active at other proteases and targets from different gene families.

It has been suggested that differences between the promiscuity index and the fragment similarity of targets might be explained by the greater chemical diversity (and hence lower fragment similarity) of ligands for promiscuous targets. To investigate this, we return to the 53 kinase profiling data set and quantify their promiscuity (since each compound has been assayed at multiple kinases, this provides a better test case than the literature/patent data). One approach simply uses the average SAR similarity of a kinase vs the remaining 52 kinases. In addition, we calculate the promiscuity indices P1, P2, and P3 described by Paolini and colleagues,[15] defining as actives those compounds having 70% inhibition or more at a given target. The average SAR similarity and Paolini's P3 are related measures of promiscuity, and the high ranking of the kinases KDR and FLT3 (first and second) by these measures is consistent with our experience with those targets (Figure 6).

Two measures for quantifying the diversity of actives were calculated: (1) the number of nonredundant actives, in which only compounds having Daylight similarity less than 0.85 remain, and (2) the average distance computed over all pairs of actives at a target. Both measures of diversity are positively correlated with the measures of promiscuity described above (i.e., more promiscuous targets have a greater diversity of actives), although only the nonredundant actives count is significantly different when comparing low vs high diversity groups with respect to the P3 measure (Figure 3) and the average SAR similarity (not shown).

In summary, we believe that targets of different gene families having high fragment similarity of actives present opportunities for cross fragment-based drug discovery; a well-developed understanding of fragment−activity relationships at one target can be applied to a target unrelated by sequence. It is not unreasonable to expect that peptidic fragments of protease ligands may be observed in ligands of peptide GPCRs, even when the overall similarity of actives (i.e., assembled fragments) is low. Another possible application of fragment driven target similarities is the identification of new pharmacology for existing compounds in the spirit of Wermuth's SOSA approach[35] and our drugs in other drugs (DIOD) concept.[36] What the present work provides is a quantitative means for identifying target pairs in which cross-utilization of fragments or entire molecules may prove fruitful.

Discussion

The decomposition of molecules into fragments allows the comparison of molecules using standard cheminformatics

Table 4. Proteins Pairs Having High Fragment Similarity and Insignificant Sequence Similarity[a]

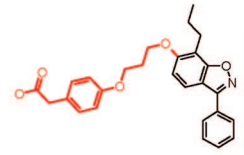
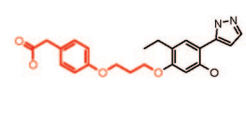
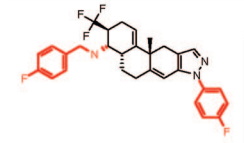
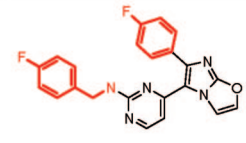
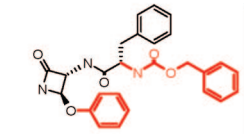
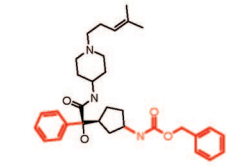
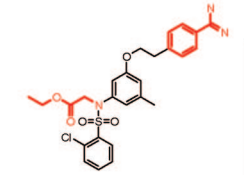
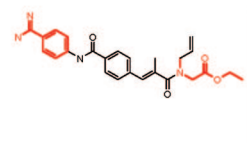
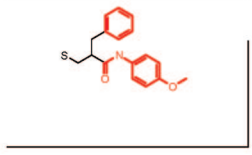
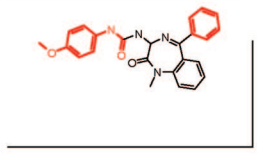
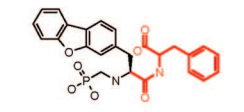
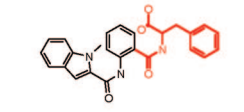
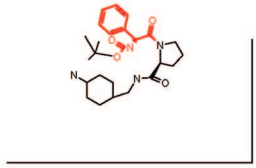
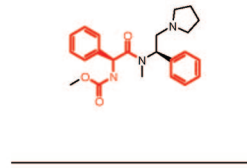
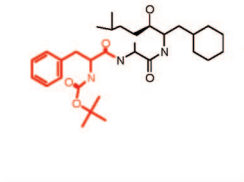
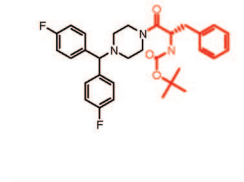
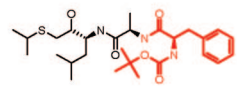
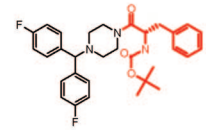
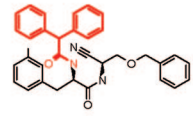
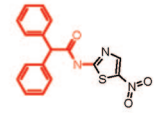
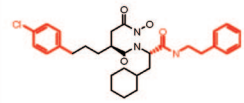
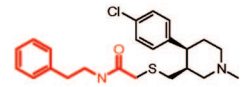
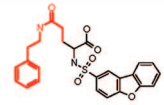
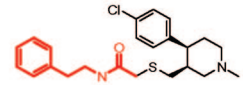
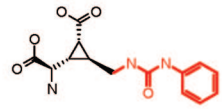
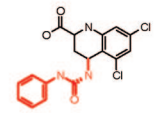
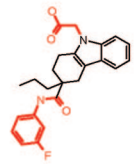
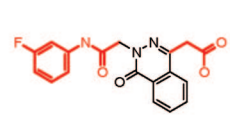
Compound 1 Target (gene family), Activity in nM, (Reference)	Compound 2 Target (gene family) Activity in nM, (Reference)	Compound 1	Compound 2
PPARG (NHR), Act=32 (Bioorg. Med. Chem. Lett. 13: 931)	LTB4R (GPCR), Act=54 (J. Med. Chem. 37: 2411)		
NR3C1 (NHR), Act=3 (WO 2004/026248 A2)	MAPK14 (Kinase), Act=90.9 (WO 2004/110990 A2)		
CTSL (Protease), Act=17 (Bioorg. Med. Chem. Lett. 13: 139)	CHRM3 (GPCR), Act=5.2 (Bioorg. Med. Chem. 8: 825)		
F2 (Protease), Act=300 (WO 98/01422 A1)	LTB4R (GPCR), Act=65 (Bioorg. Med. Chem. 5: 971)		
MME (Protease), Act=26 (Bioorg. Med. Chem. 6: 441)	CCKAR (GPCR), Act=165 (J. Med. Chem. 36: 4276)		
MME (Protease), Act=12 (J. Med. Chem. 43: 488)	CCKAR (GPCR), Act=350 (Eur. J. Med. Chem. 39: 85)		
F2 (Protease), Act=6.8 (US 5714485 A)	OPRK1 (GPCR), Act=178 (Bioorg. Med. Chem. Lett. 15: 1279)		
REN (Protease), Act=200 (J. Med. Chem. 31: 2264)	CACNA1B (Ion channel), Act=250 (WO 99/55688 A1)		

Table 4. Continued

Compound 1 Target (gene family), Activity in nM, (Reference)	Compound 2 Target (gene family) Activity in nM, (Reference)	Compound 1	Compound 2
REN (Protease), Act=700 (J. Med. Chem. 31: 532)	CACNA1B (Ion channel), Act=250 (WO 99/55688 A1)		
CTSB (Protease), Act=10.2 (J. Med. Chem. 44: 4524)	CACNA1B (Ion channel), Act=250 (US 20030199523 A1)		
MMP3 (Protease), Act=7.31 (J. Med. Chem. 44: 2333)	SLC6A3 (Transporter), Act=23 (J. Med. Chem. 48: 7970)		
MMP3 (Protease), Act=325 (WO 98/09957 A1)	SLC6A3 (Transporter), Act=112 (J. Med. Chem. 48: 7970)		
GRM2 (GPCR), Act=530 (Bioorg. Med. Chem. 13: 6556)	GRIN1 (Ion channel), Act=2 (J. Med. Chem. 37: 3956)		
GPR44 (GPCR), Act=22 (WO 2006/070325 A3)	AKR1B1 (Other), Act=22 (J. Med. Chem. 35: 457)		

[a] Official gene symbols are defined as follows: AKR1B1, aldose reductase; CACNA1B, voltage-gated calcium channel; CCKAR, cholecystokinin A receptor; CHRM3, M3 muscarinic receptor; CTSB, cathepsin B; CTSL, cathepsin L1; DRD1, dopamine receptor D1; F2, thrombin; GPR44, G-protein-coupled receptor 44; GRIN1, NMDA receptor 1; GRM2, glutamate receptor, metabotropic 2; HTR2A, serotonin receptor 2A; LTB4R, leukotriene B4 receptor; MAPK14, p38 MAP kinase; MME, membrane metalloendopeptidase; MMP3, matrix metalloproteinase 3; NR3C1, glucocorticoid receptor; OPRK1, opioid receptor κ1; PPARG, PPAR γ; REN, renin; SLC6A2, noradrenaline transporter; SLC6A3, dopamine transporter; SLC6A4, serotonin transporter. IC_{50} or K_i activities expressed in nM for representative compounds. References for journals indicate the volume number and first page; otherwise, patent numbers are indicated.

applications. We have demonstrated their use for building highly interpretable and mathematically simple naive Bayes models of activity in single concentration kinase binding assays. Retrospective cross-validation and prospective testing of compounds predicted active indicate the usefulness of fragment fingerprints in predictive modeling. In contrast to some popular fingerprint methods, the connection to chemical structure is obvious. This facilitates the conversion of variable importance (e.g., the probability of observing activity given the presence of a given fragment) into ideas for synthesizing novel molecules.

In addition, the comparison of fragment distributions in ligands for two proteins provides a means of assessing protein similarity. The fragment-derived similarities complement sequenced-derived similarities, since they can establish relationships between proteins modulated by small molecules with similar fragment composition, even in cases where they exhibit weak or absent sequence similarity. Fragment-derived similarities require only a moderate number of actives at each target, thus extending their applicability to protein pairs where insufficient data are available for directly calculating their SAR similarity using compounds assayed at both targets. The identification of protein pairs having high fragment similarity is useful for initiating screening efforts at the second target, using a possibly greater number and diversity of actives at the

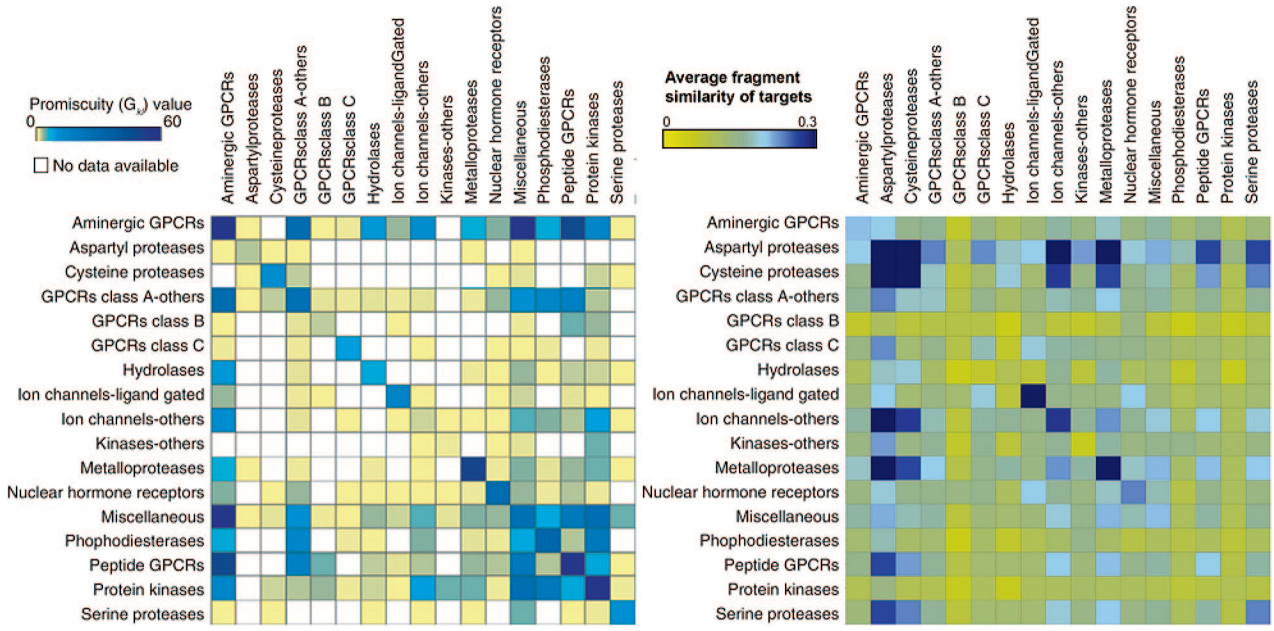


Figure 5. Comparison of the promiscuity index[15] determined using proprietary and publicly available SAR data, within and between various protein classes (left), and average fragment similarity of targets determined using journal and patent data, within and between the same classes (right). Two target classes (enzymes, others; oxidoreductases) were removed because of inadequate coverage in our database of public actives or uncertainty in how the Paolini terms relate to Proteome[33] annotations generally employed in this work. Adapted by permission from Macmillan Publishers Ltd: *Nature Biotechnology*, ref 15, Copyright 2006.

first, and for understanding the potential for cross-activity of compounds beyond that enabled by sequence comparisons.

Methods

Decomposition of Small Molecules into Fragments Using Dicer. The fragmentation scheme applied in this work shares underlying principles with RECAP.[2] Like RECAP, Dicer will not cleave a ring bond. Whereas RECAP contains a list of 11 bond types that can be broken, dicer takes a more expansive view and breaks the bond types indicated in Figure 1.

These bond breaking patterns were identified by chemists as being likely regions of interest in molecule synthesis. Unlike RECAP, Dicer is fully recursive and so can generate fragments of fragments and overlapping fragments. When limiting the size of fragments to between 4 and 17 heavy atoms, RECAP rules give fewer fragments (~40% when compared to Dicer fragments) with slightly lower molecular weight and ClogP (within 10% of Dicer fragments). The maximum number of bonds to be simultaneously broken is a user controlled parameter (set to 3 in this work); this restriction prevents combinatorial explosions, which produce minimally interesting fragments. Only fragments containing between 4 and 17 heavy atoms are used in this work; limiting the size of fragments keeps the molecular weight of reassembled molecules within a reasonable range.

Defining a Fragment Fingerprint. The total number of fragments occurring in our 151 000 compound data set of journal and patent actives is 226 453. When a threshold of 1 occurrence per 1000 molecules is imposed, the number of fragments used for representing compounds or targets is reduced to 1126. A similar reduction for the kinase-assayed compounds results in 1815 fragments used for representing compounds and targets.

Ignoring rare fragments results in smaller fingerprints and increases the speed of calculations. It is possible to use smaller values, but 1 occurrence per 1000 molecules is a very rare fragment. For calculating the fragment similarity of two proteins, including rarer fragments has no impact since their effect is proportional to the rate of occurrence (see formula below). For naive Bayes models, the use of Dirichlet priors and other "smoothing" transformations reduces the effect of rare features. Naive Bayes models using a minimum occurrence of 1 fragment per 10 000 molecules give results numerically identical to those in Table 1.

Fragment and SAR Similarity. The fragment similarity of two proteins, which ranges from 0 to 1, is obtained by calculating the frequency of occurrence of each fragment in active molecules ((number of occurrences of fragment)/(number of actives molecules)). For two targets described by the frequencies of N fragments, the Tanimoto similarity is calculated using

$$\mathrm{FragSim}_{1,2} = \frac{\sum_{k=1}^{N_{\mathrm{frags}}} \mathrm{Fr}_{1,k}\mathrm{Fr}_{2,k}}{\sum_{k=1}^{N_{\mathrm{frags}}} \mathrm{Fr}_{1,k}^{\,2} + \mathrm{Fr}_{2,k}^{\,2} - \mathrm{Fr}_{1,k}\mathrm{Fr}_{2,k}}$$

where $\mathrm{Fr}_{1,k}$ is the frequency of the kth fragment among actives of protein 1. The calculation is exemplified in Figure 3. SAR similarity can be calculated in a similar manner, using the activity of N compounds assayed at each receptor:

$$\mathrm{SARSim}_{1,2} = \frac{\sum_{k=1}^{N_{\mathrm{cpds}}} \mathrm{Act}_{1,k}\mathrm{Act}_{2,k}}{\sum_{k=1}^{N_{\mathrm{cpds}}} \mathrm{Act}_{1,k}^{\,2} + \mathrm{Act}_{2,k}^{\,2} - \mathrm{Act}_{1,k}\mathrm{Act}_{2,k}}$$

where $\mathrm{Act}_{1,k}$ is a value between 0 and 3 denoting the activity range of compound k at protein 1. Each pair of kinases has at least 570 compounds assayed in common, with an average of 7875.

In order to relate single-concentration inhibition (20 μM) to the probability of obtaining an IC_{50} value less than 1 μM in a concentration–response curve, percent inhibition values are binned into four groups for SAR similarity calculations: 0–50% = 0 (negligible), 50–70% = 1 (low), 70–90% = 2 (moderate), 90% or more = 3 (high). We have found empirically that this scheme represents reasonably well the probability of having $IC_{50} \leq 1$ μM, indicated in parentheses for each group.

Naive Bayes Models. For a one fragment model, Bayes rule can be used to calculate the probability of observing activity ($A = 0$ for inactives, $A = 1$ for actives) given the presence of fragment i:

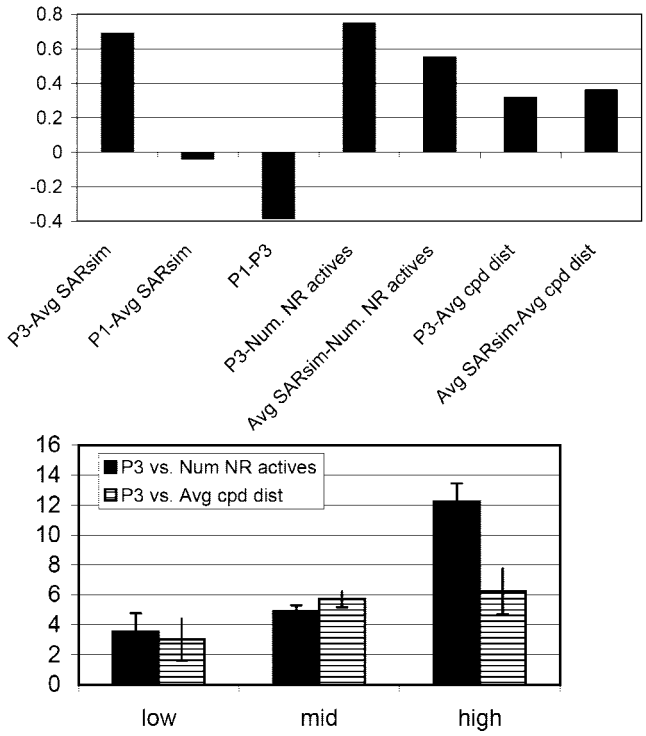


Figure 6. (Top) Comparison of kinase promiscuity as measured by the average SAR similarity of a kinase vs the other 52 kinases and the measures P1, P2, and P3 defined by Paolini and co-workers.[15] Actives are defined as having ≥70%inhibition at 20 μM test concentration. The measure P1 is inconsistent with average SAR similarity, and the low ranking of FLT3 (3rd least promiscuous) and KDR (14th least promiscuous) is inconsistent with our experience. The measure P2 is not useful in this context, as all pairs of kinases have at least one active in common and all kinases have the same promiscuity (P2 = 52). In addition, correlations between promiscuity metrics and two measures of actives diversity are shown: the number of nonredundant actives (i.e., only retaining actives that have Daylight similarity less than 0.85 compared to every other active) and the average Daylight distance (1 − similarity) between actives at each target. (Bottom) Average P3 promiscuity when grouping kinases into the lowest 10%, highest 10%, and middle 80% for each chemical diversity measure. Low−high differences are not statistically significant at 95% confidence when using the average compound distance chemical diversity measure.

$$P(A=1|\mathrm{Fr}_i) = P(A=1)\cdot P(\mathrm{Fr}_i|A=1)/P(\mathrm{Fr}_i)$$

$P(A=1)$ is simply the hit rate from the assay, and $P(\mathrm{Fr}_i)$ is the probability of observing fragment i. Generalizing to N fragments and assuming conditional independence of fragments,

$$P(\mathrm{Fr}_i|A,\mathrm{Fr}_j) = P(\mathrm{Fr}_i|A)$$

one can show that the log likelihood ratio equals

$$\mathrm{LLR} = \ln\frac{P(A=1|\mathrm{Fr}_1,\mathrm{Fr}_2,...,\mathrm{Fr}_N)}{P(A=0|\mathrm{Fr}_1,\mathrm{Fr}_2,...,\mathrm{Fr}_N)} = \ln\frac{P(A=1)}{P(A=0)} + \sum_{i=1}^{N}\ln\frac{P(\mathrm{Fr}_i|A=1)}{P(\mathrm{Fr}_i|A=0)}$$

For a two-class problem (active vs inactive), the log likelihood ratio is equal to

$$\mathrm{LLR} = \ln\frac{P(A=1|\mathrm{Fr}_1,\mathrm{Fr}_2,...,\mathrm{Fr}_N)}{P(A=0|\mathrm{Fr}_1,\mathrm{Fr}_2,...,\mathrm{Fr}_N)} = \ln\frac{P(A=1|\mathrm{Fr}_1,\mathrm{Fr}_2,...,\mathrm{Fr}_N)}{1-P(A=1|\mathrm{Fr}_1,\mathrm{Fr}_2,...,\mathrm{Fr}_N)}$$

allowing the probability of activity given the presence/absence of fragments Fr_i in a compound to be calculated as

$$P(A=1|\mathrm{Fr}_1,\mathrm{Fr}_2,...,\mathrm{Fr}_N) = \frac{\exp(\mathrm{LLR})}{1+\exp(\mathrm{LLR})}$$

The standard Dirichlet prior is assumed on Fr_i, to give smooth probability estimates:

$$P(\mathrm{Fr}_i|A=1) = (\text{number of actives containing } \mathrm{Fr}_i + 1)/(\text{number of actives} + 2)$$

Positive Predictive Value Calculation. Positive predictive values (PPV) are calculated from the sensitivity (TP), specificity (TN), and prevalence Prev from

$$\mathrm{PPV} = \frac{\mathrm{TP}\cdot\mathrm{Pr}ev}{\mathrm{TP}\cdot\mathrm{Pr}ev + (1-\mathrm{TN})(1-\mathrm{Pr}ev)}$$

PPV ranges from 0 (no precision) to 1 (perfect precision). As TN tends to 1, PPV tends to 1; in other words, even a model correctly classifying only 1% of actives will have perfect precision, since it will never incorrectly classify an inactive compound.

Biochemical Assays. Recombinant protein (5–10 mU) is incubated with 0.2 mM EDTA, 8 mM MOPS having pH 7.0, 10 mM magnesium acetate, 50 μM substrate, 20 μM concentration of test substance, and γ-^{33}P-ATP, having concentration equal to K_m for the enzyme. The final reaction volume is 25 μL. Addition of MgATP initiates the reaction, which is followed by a 40 min incubation at room temperature. The reaction is stopped by the addition of 5 μL of a 3% phosphoric acid solution. A 10 μL sample is plotted onto a P30 filtermat and washed three times for 5 min in 75 mM phosphoric acid and once in methanol before drying and scintillation counting.

Acknowledgment. We thank Jibo Wang for assistance in figure preparation, and John Toth, Mary Mader, and Jon Erickson for helpful discussions on kinase cheminformatics. We thank members of the Global Computational Chemistry Group for suggestions and review of this manuscript.

Supporting Information Available: An Excel spreadsheet listing the most frequently occurring fragments in journal/patent actives for each gene family in Table 3. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Hanessian, S. Man, machine and visual imagery in strategic synthesis planning: computer-perceived precursors for drug candidates. *Curr. Opin. Drug Discovery Dev.* **2005**, *8*, 798–819.
(2) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP—retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
(3) Bemis, G. W.; Murcko, M. A. Properties of known drugs. 2. Side chains. *J. Med. Chem.* **1999**, *42*, 5095–5099.
(4) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
(5) Berk, S. C.; Rohrer, S. P.; Degrado, S. J.; Birzin, E. T.; Mosley, R. T.; Hutchins, S. M.; Pasternak, A.; Schaeffer, J. M.; Underwood, D. J.; Chapman, K. T. A combinatorial approach toward the discovery of non-peptide, subtype-selective somatostatin receptor ligands. *J. Comb. Chem.* **1999**, *1*, 388–396.
(6) Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S.; Chang, R. S. L.; Lotti, V. J.; Cerino, D. J.; Chen, T. B.; Kling, P. J.; Kunkel, K. A.; Springer, J. P.; Hirshfield, J. Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* **1988**, *31*, 2235–2246.
(7) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.
(8) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are target-family-privileged substructures truly privileged? *J. Med. Chem.* **2006**, *49*, 2000–2009.
(9) Muller, G. Medicinal chemistry of target family-directed masterkeys. *Drug Discovery Today* **2003**, *8*, 681–691.
(10) Erlanson, D. A. Fragment-based lead discovery: a chemical update. *Curr. Opin. Biotechnol.* **2006**, *17*, 643–652.

(11) Zartler, E. R.; Shapiro, M. J. Fragonomics: fragment-based drug discovery. *Curr. Opin. Chem. Biol.* **2005**, *2005*, 366370.
(12) Vieth, M.; Higgs, R. E.; Robertson, D. H.; Shapiro, M.; Gragg, E. A.; Hemmerle, H. Kinomics—structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. Biophys. Acta* **2004**, *1697*, 243–257.
(13) Vieth, M.; Sutherland, J. J.; Robertson, D. H.; Campbell, R. M. Kinomics: characterizing the therapeutically validated kinase space. *Drug Discovery Today* **2005**, *10*, 839–846.
(14) Frye, S. V. Structure−activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era. *Chem. Biol.* **1999**, *6*, R3–R7.
(15) Paolini, G. V.; Shapland, R. H.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24*, 805–815.
(16) Izrailev, S.; Farnum, M. A. Enzyme classification by ligand binding. *Proteins* **2004**, *57*, 711–724.
(17) Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles, J. H.; Davies, J. W. "Bayes affinity fingerprints" improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multitarget drugs a feasible concept? *J. Chem. Inf. Model.* **2006**, *46*, 2445–2456.
(18) Givehchi, A.; Bender, A.; Glen, R. C. Analysis of activity space by fragment fingerprints, 2D descriptors, and multitarget dependent transformation of 2D descriptors. *J. Chem. Inf. Model.* **2006**, *46*, 1078–1083.
(19) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206.
(20) Swindells, M. B.; Overington, J. P. Prioritizing the proteome: identifying pharmaceutically relevant targets. *Drug Discovery Today* **2002**, *7*, 516–521.
(21) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure−activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
(22) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82–85.
(23) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903–911.
(24) Seel, M.; Turner, D. B.; Willett, P. Effect of parameter variations on the effectiveness of HQSAR analyses. *Quant. Struct.−Act. Relat.* **1999**, *18*, 245–252.
(25) Instead of using binary fingerprints, it is possible to represent the number of occurrences of features. We explored this possibility for calculating the fragment similarity of two proteins, which we did before the naive Bayes work. This resulted in no difference for fragments and only a small improvement of 0.01 in the correlation coefficient (*r*) for MACCS keys. For this reason, we did not apply counted fingerprints in naive Bayes models.
(26) We expect similar results if K_i or IC_{50} data were used for the same compounds. However, we do not have K_i/IC_{50} data for every active compound from the single concentration assay.
(27) While enrichments at 0.1% might be more appropriate for large screening sets exceeding 1 million compounds, the relatively small data set used here would prohibit meaningful analysis at this threshold.
(28) Stouch, T. R.; Kenyon, J. R.; Johnson, S. R.; Chen, X. Q.; Doweyko, A.; Li, Y. In silico ADME/Tox: why models fail. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 83–92.
(29) Fabian, M. A.; Biggs, W. H., 3rd; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D.; Benedetti, M. G.; Carter, T. A.; Ciceri, P.; Edeen, P. T.; Floyd, M.; Ford, J. M.; Galvin, M.; Gerlach, J. L.; Grotzfeld, R. M.; Herrgard, S.; Insko, D. E.; Insko, M. A.; Lai, A. G.; Lelias, J. M.; Mehta, S. A.; Milanov, Z. V.; Velasco, A. M.; Wodicka, L. M.; Patel, H. K.; Zarrinkar, P. P.; Lockhart, D. J. A small molecule−kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329–336.
(30) Chen, J.; Zhang, X.; Fernandez, A. Molecular basis for specificity in the druggable kinome: sequence-based analysis. *Bioinformatics* **2007**, *23*, 563–572.
(31) Fernandez, A.; Maddipati, S. A priori inference of cross reactivity for drug-targeted kinases. *J. Med. Chem.* **2006**, *49*, 3092–3100.
(32) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): theory and application. *J. Chem. Inf. Model.* **2007**, *47*, 279–294.
(33) Hodges, P. E.; Carrico, P. M.; Hogan, J. D.; O'Neill, K. E.; Owen, J. J.; Mangan, M.; Davis, B. P.; Brooks, J. E.; Garrels, J. I. Annotating the human proteome: the Human Proteome Survey Database (HumanPSD) and an in-depth target database for G protein-coupled receptors (GPCR-PD) from Incyte Genomics. *Nucleic Acids Res.* **2002**, *30*, 137–141.
(34) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
(35) Wermuth, C. G. Selective optimization of side activities: the SOSA approach. *Drug Discovery Today* **2006**, *11*, 160–164.
(36) Siegel, M. G.; Vieth, M. Drugs in other drugs: a new look at drugs as fragments. *Drug Discovery Today* **2007**, *12*, 71–79.

JM701399F