QSAR Models for Predicting the Similarity in Binding Profiles for Pairs of Protein Kinases and the Variation of Models between Experimental Data Sets

Robert P. Sheridan,* Kiyean Nam, Vladimir N. Maiorov, Daniel R. McMasters, and Wendy D. Cornell

Chemistry Modeling and Informatics Department, Merck Research Laboratories, RY50SW-100, Rahway, New Jersey 07065

Received May 14, 2009

We propose a direct QSAR methodology to predict how similar the inhibitor-binding profiles of two protein kinases are likely to be, based on the properties of the residues surrounding the ATP-binding site. We produce a random forest model for each of five data sets (one in-house, four from the literature) where multiple compounds are tested on many kinases. Each model is self-consistent by cross-validation, and all models point to only a few residues in the active site controlling the binding profiles. While all models include the "gatekeeper" as one of the important residues, consistent with previous literature, some models suggest other residues as being more important. We apply each model to predict the similarity in binding profile to all pairs in a set of 411 kinases from the human genome and get very different predictions from each model. This turns out not to be an issue with model-building but with the fact that the experimental data sets disagree about which kinases are similar to which others. It is possible to build a model combining all the data from the five data sets that is reasonably self-consistent but not surprisingly, given the disagreement between data sets, less self-consistent than the individual models.

INTRODUCTION

The protein kinase family is defined by sharing a common catalytic domain, which phosphorylates substrate peptides by transferring the γ-phosphate of ATP to the hydroxyl group of serine, threonine, or tyrosine. Protein kinases constitute one of the largest protein families with 518 potential protein kinases identified from the human genome.[1] In recent years, many diseases, such as cancer, diabetes, and inflammatory disorders have been attributed to alteration in the kinase-mediated signaling pathways. Consequently, an increasing number of protein kinases have been selected as therapeutic targets. The same involvement of protein kinases in multiple pathways that makes them attractive as drug targets also gives rise to the problem that any molecule that will bind tightly to the target kinase is also likely to bind to other kinases in other pathways and give rise to unintended side-effects. This is especially likely because most kinase inhibitors so far have been designed to bind in the well-conserved ATP binding site. Therefore, when researchers develop kinase inhibitors, they must also pay close attention to selectivity: presumably, the ideal candidate inhibitor will bind tightly to the target kinases but not to many others. The literature examining and explaining the selectivity of inhibitors for kinases and vice versa is very large. One can divide the experimental literature into those papers that test a number of compounds against a variety of kinases (for example, refs 2–6) and those that examine the structures of kinases and their cocrystallized ligands and, using this information, suggest more selective compounds for a specific kinase or monitor the changes in selectivity as specific residues are mutated (for example, refs 7–19). Molecular modeling has been applied to the selectivity problem in a variety of ways (for example, refs 20–32). One should also note the effort to quantitate the selectivity of kinase inhibitors.[33]

One of the most common approaches to address the selectivity issue in drug development is to screen kinase inhibitors against off-target kinases. Counter screening against the entire human kinome is currently impractical, but counter screening against a subset of the human kinome is a common practice in both early and late stages of drug development. Subsets are sometimes selected based on sequence homology in the catalytic domain. However, it is well established that the selectivity profile of kinases has little correlation with the overall homology of their catalytic domain sequences.[34,35] These findings present difficulties in effectively addressing selectivity of a target since it is hard to predict from their sequences which other kinases are likely to resemble the target in the types of molecules they bind.

Given that only a fraction of the kinome has been tested experimentally, it would be very useful to have a model to predict how similar any two kinases in the kinome are likely to be in the types of molecules they bind. Using such a model, one could construct a subset of kinases that are most likely to share a similar inhibition profile with the target kinase. As an alternative application, one can assess "druggability" of a given kinase by estimating the potential difficulty of achieving selectivity: Some kinases are similar to many other kinases; others are similar to only a few. Finally, the model can be used to target-hop, that is, to find inhibitors for a new kinase by testing inhibitors from kinases that are predicted to be similar.

Here, we develop QSAR models calibrated against data sets of multiple compounds tested on multiple kinases. The

* Corresponding author e-mail: sheridan@merck.com; telephone: 732-594-3859.

10.1021/ci900176y CCC: $40.75
Published on Web 07/29/2009

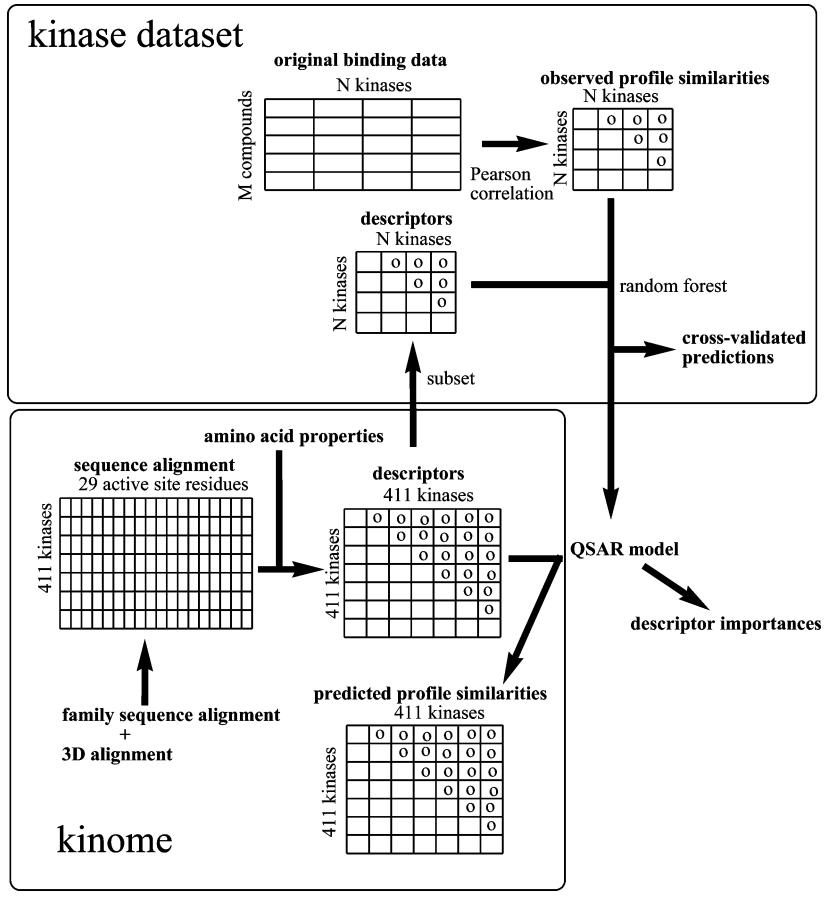

Figure 1. Schematic diagram of workflow.

models determine the relative importance of residues in the ATP binding site in terms of the contributions of the residues to the given kinase inhibition profile, and we can use this information to predict similarity between two untested kinases. We use the models to predict inhibition-profile-based interkinase relationships for 80% of the human kinome (411 kinases out of 518). Unfortunately, the models made from the different data sets disagree in their predictions, and we can show that this is likely because the experimental determinations of which kinases are similar to which others can vary strongly between data sets.

METHODS

Figure 1 shows an overview of the workflow described in this section. There are several steps:

1. Generate an alignment of active site residues for most of the kinome.
2. For each pair of kinases in the kinome, generate descriptors summarizing the difference between the amino acids at specific active site positions.
3. For each pair of kinases in each experimental data set, generate a similarity of their binding profiles.
4. For each data set calculate and validate a QSAR model.
5. Using the QSAR models, predict the binding profile similarity for each pair of kinases in the kinome.

Kinase Nomenclature. There are multiple nomenclature systems for kinases. We use the kinase nomenclature based on the work of Manning et al.[1] who first presented the human kinome. This has a supporting Web site http://kinase.com/human/kinome/phylogeny.html. These workers divide the kinome into eight groups based on sequence: AGC, CAMK, CK1, CMGC, STE, TK, TKL, and Other.

Alignment of Active Site Residues. We will use the term "binding profile" to refer to how any given kinase binds a set of common compounds. We make the assumption that the residues near the ATP-binding site provide almost all the contributions to the binding profile and therefore are restricting the analysis to a set of 29 residues immediately surrounding that site. The correspondence of our active site residue numbering to PKA kinase numbering is shown in Table 1. Finding which residues correspond to the 29 active site residues for all the kinases in the kinome is difficult since it is very hard to find a credible global sequence alignment for all the groups, which can be quite diverse. We therefore did the alignment in three steps:

1. Multiple sequence alignments of human kinase sequences within a group, which are reliable, were taken from http://kinase.com/human/kinome/phylogeny.html.
2. The correspondence between active site residues between groups was taken from a 3D superposition of kinases for which a crystal structure was available in the Protein Data Bank.[36] These are listed in Table 2. At least one member of each group is represented. The structures were overlaid onto 1AD5 with the ICM program (version 3.4),[37] using α carbon coordinates and no sequence information. The rmsd values of the coordinates of aligned α carbons were in the range of 1.2−2.4 Å for 177−240 aligned residues across all the structures. After visual inspection, we concluded that the derived overlays are consistent with each other and can be used for further derivation of a 29-residue alignment based on spatial proximity of α carbons. For instance, we may make a reasonable guess of the correspondence between active site residues in 1CM8 (P38 from the CMGC group) with active site

Table 1. Correspondence of the Numbering of Residues in the Active Site Using Protein Kinase A As the Standard

residue number	PKA numbering	notation
1	L49	Gly-rich loop
2	G50	Gly-rich loop
3	T51	Gly-rich loop
4	G52	Gly-rich loop
5	S53	Gly-rich loop
6	F54	Gly-rich loop
7	G55	Gly-rich loop
8	R56	Gly-rich loop
9	V57	Gly-rich loop
10	A70	
11	K72	
12	E91	
13	L95	
14	V104	
15	M118	
16	M120	gatekeeper
17	E121	hinge
18	Y122	hinge
19	V123	hinge
20	P124	hinge
21	G125	hinge
22	G126	hinge
23	S130	
24	E170	
25	N171	
26	L173	
27	T183	
28	D184	DFG-in/out
29	F187	

residues in, say, 1APM (cAMP-dependent protein kinase) from the AGC group.

3. Since one has identified the active site residues in a key member of each group, one has also identified the residues in all members via the multiple sequence alignments for that group in step 1.

Not all members of the kinome are similar enough in sequence to a protein of known 3D structure to be assigned in this way. On the basis of the protein kinase crystal structures available at the time, we were able to make assignments for 411 out of 518 kinases in the human kinome. The active site residue alignments are in Supporting Information. Figure 2 shows the placement of these 29 residues in the active site of human PKA.

QSAR Descriptors from the Alignment. We are representing the similarity of a pair of kinases based on the properties of the 29 residues in the active site. We considered 15 physical properties of amino acids. Hence, there will be $29 \times 15 = 435$ descriptors in the QSAR model (plus 29 more as will be shown later). The list of residue properties we used is in Supporting Information. Note that some are different measures of the same property from different authors. The fact that many of the descriptors are highly correlated is not a problem, as will be discussed below.

Let P(p,k) be the value of property p of amino acid type k. One can normalize P by the mean and standard deviation for the values over all k. We define P'(p,k) as the z-score of property p of amino acid k relative to all 20 amino acids. For example for p = "bulkiness", the mean and standard deviation over all amino acids is 13.4 ± 6.7. Hence for p = "bulkiness" and k="W" (tryptophan), P(p,k)=21.67 and P'(p,k)=(21.67−13.4)/6.7=1.2.

Imagine a pair of kinases A and B. The descriptor associated with position i (out of 29) and property p is:

$$d(i,p) = |P'(p,k_{Ai}) - P'(p,k_{Bi})|$$

where k_{Ai} is the amino acid type in position i for kinase A. The more A and B are alike at that property at that position, the closer $d(i,p)$ is to zero.

We also include at each position an extra "amino acid identity" descriptor $d(i,x) = 1$ if $k_{Ai} = k_{Bi}$, but $d(i,x) = 0$ otherwise. Adding these 29 "identity" descriptors to the 435 described above, we have a total of 464 descriptors.

Kinase Data Sets. We used five sets of data, summarized in Table 3, to construct models. One complication is that literature data sets always use a nomenclature system other than that of Manning et al., and it was necessary for us to convert the name of the kinases in each data set "by hand". Not all the kinases from any particular data set could be used. Sometimes the conversion in nomenclature could not be made with confidence, or the corresponding sequence was one that we could not sensibly align.

1. Dundee. Percent inhibition data was produced by the Dundee Division of Signal Transduction Therapy Consortium for a total of 950 Merck molecules tested on 51 distinct kinases. Note that not all the kinases in the Dundee panel are from a human source. However, in those cases, the active site residues are identical, and we can use the alignment for the nearest human kinase. Details on the Dundee assay system can be found in Bain et al.[2] Many elements of the Dundee table are missing because this data was not collected as part of a systematic study. Some pairs of kinases have up to ~900 compounds tested in common, some have very few or none, so we were not able to include every pair in our model (see below).

2. Karaman et al. K_d values (in nM) for 60 nonproprietary compounds tested on 317 human kinases (including mutants) were taken from Karaman et al.[5] For our purposes, the activities are converted to $-\log(K_d)$. We have an alignment for only 160 unique kinases from Karaman et al., so only those could be included in our QSAR model.

3. Bamborough et al. These investigators[3] measured percent control activity for 577 compounds over 203 kinases. We obtained the complete data matrix, with the molecules anonymized, directly from the authors. We are able to assign alignments for 191 of the kinases.

4. Federov et al. These authors[4] measured shifts in thermostabilities for 156 compounds on 60 Ser/Thr kinases. This is a surrogate for measurements on binding, where a shift of 4 Celsius is considered roughly equivalent to a binding affinity of 1 μM. We can assign alignments to 49 of the kinases.

5. Melnick et al. These authors[6] measured IC_{50} values for 1400 compounds against 35 Tyr kinases in a cell-based assay. Results are reported in μM, and we used $-\log(IC_{50})$ as the activity. We can assign an alignment of 28 of the kinases.

The diversity of active site sequences in each data set as measured by the distribution of pairwise sequence distances using the PROTDIST module of PHYLIP (http://evolution.genetics.washington.edu/phylip.html) shows that most of the data sets are similarly diverse, with the exception of Melnick et al., which is much less diverse. This is not surprising given that it has the fewest sequences and it is confined to Tyr-

Table 2. Kinase Crystal Structures used in a Derivation of the 29 Residue Alignment

PDB code	resolution	PKR id	name	species	group	nearest human kinase
1APM	2.0	968	PKA	mouse	AGC	PKA__PKACa
1CDK	2.0	964	PKA(alpha)	rabbit	AGC	PKA__PKACa
1FOT	2.8	972	TPK1	yeast	AGC	PKA__PKACb
1O6K	1.7	1932	Akt2/PKBbeta	human	AGC	AKT__AKT2
1OMW	2.5	1953	BARK1	bovine	AGC	GRK_BARK_BARK1
1TKI	2.0	16836	TITN	human	CAMK	MLCK__TTN
1IA8	1.7	1688	CHK1	human	CAMK	CAMKL_CHK1_CHK1
1KOB	2.3	17036	Twitchin	sea hare	CAMK	MLCK__SgK085
1KWP	2.8	671	MAPKAPK2	human	CAMK	MAPKAPK_MAPKAPK_MAPKAPK2
1CKI	2.3	1373	KC1D	rat	CK1	CK1__CK1d
1CSN	2.0	1681	CK1	yeast	CK1	CK1__CK1g2
1BLX	1.9	1832	CDK6	human	CMGC	CDK2_CDK4_CDK6
1P38	2.1	660	MK14	mouse	CMGC	MAPK_p38_p38a
1CM8	2.4	651	P38gamma	human	CMGC	MAPK_p38_p38g
1GNG	2.6	1644	GSK3b	human	CMGC	GSK_GSK3B
1H1S	2.0	1775	CDK2	human	CMGC	CDK_CDC2_CDK2
1H4L	2.65	1827	CDK5	human	CMGC	CDK_CDK5_CDK5
1JNK	2.3	645	JNK3	human	CMGC	MAPK_JNK_JNK3
1PME	2.0	527	ERK2	human	CMGC	MAPK_ERK_ERK2
1MUO	2.9	74	AurA/Aur2	human	Other	Other_AUR_AurC
1F3M	2.3	409	PAk1	human	STE	STE20_PAKA_PAK1
1AD5	2.6	1540	HCK	human	TK	Src__HCK
1FGK	2.0	1797	FGFR1	human	TK	FGFR__FGFR1
1GJO	2.4	1802	FGFR2	human	TK	FGFR__FGFR2
1FVR	2.2	204	TIE2	human	TK	Tie__TIE2
1IR3	1.9	1006	INSR	human	TK	InsR__INSR
1MP8	1.6	1101	FAK	human	TK	Fak__FAK
1MQB	2.3	1225	EphA2	human	TK	Eph__EphA2
1JPA	1.91	1713	EphB2	mouse	TK	Eph__EphB2
1K2P	2.1	1563	BTK	human	TK	Tec__BTK
1K9A	2.5	1951	CSK	rat	TK	Csk_CSK
1LUF	2.05	857	Musk	rat	TK	Musk__MUSK
1M17	2.6	1941	EGFR	human	TK	EGFR__EGFR
1OPK	1.8	1814	ABL1	mouse	TK	Abl__ABL
1P4O	1.5	1511	IGF1R	human	TK	InsR__IGF1R
1QPC	1.6	877	P56-LCK	human	TK	Src__LCK
1R0P	1.8	999	MET	human	TK	Met__MET
1SM2	2.3	1255	ITK	human	TK	tec__ITK
1VR2	2.4	1307	KSG2_ARATH	human	TK	VEGFR_KDR
2PTK	2.35	388	p60-Src	chicken	TK	src__SRC
1B6C	2.6	449	TGFbetaR1	human	TKL	STKR_Type1_TGFbR1

kinases. The variation of the sequences at each position for the data sets is shown in Figure 3.

QSAR End Point: Measure of the "Profile Similarity" of Kinase Pairs. For any QSAR study we need a set of descriptors (see above paragraphs) and a set of activities. In this paper we will treat each unique pair of kinases (not including a kinase paired with itself) as a unit in the QSAR and the "activity" for the pair will be the similarity in binding profile between the two kinases. The naming convention for the pair is that the kinases will be in alphabetical order (in the Manning et al. system) separated by "@". Our QSAR models will be able to predict the pairwise similarity of any two kinases, but cannot say anything about the binding of a particular molecule on a particular kinase.

The binding profile similarity for a pair of kinases will be the Pearson correlation coefficient between the activities (e.g., $-\log(K_d)$ or percent inhibition) of a set of molecules that were tested on both kinases. In principle, the correlation can range from 1 to -1, but in practice the range is slightly lower than 1 to a small negative value like -0.3. In some ways the Pearson correlation coefficient is a very crude metric of how much two kinases are "alike". One can easily imagine more sophisticated metrics, but the Pearson correlation has the advantage that one need not know the chemical structure of the compounds. This is important because our experience has been that many kinase data sets in the literature contain proprietary compounds and that authors are more willing to release them if the compounds are not identified. An example of kinase pairs with high and low Pearson correlations are shown in Figure 4.

Some data sets have complications in regards to taking the Pearson correlation coefficient. Since much data is missing in the Dundee set, we had to make a compromise between including more kinases pairs in the model at the expense of having fewer compounds tested in common for those pairs. Here, we considered only those pairs of kinases that had ~200 or more compounds tested in common. We compared the Pearson correlation for kinase pairs that had ~900 compounds in common to the correlation on the same kinase pairs when we selected ~200 out of 900 compounds at random and found them to be fairly close. That is, the Pearson correlation coefficient is not sensitive to the number of compounds, at least when the number is >200.

In the Karaman et al. data set, there are several kinases on which all molecules except staurosporine are "inactive" at the criterion of $-\log(K_d) = -4.0$. The consequence is that many kinase pairs will appear to have correlations of 1.0 because a linear regression is being done on effectively

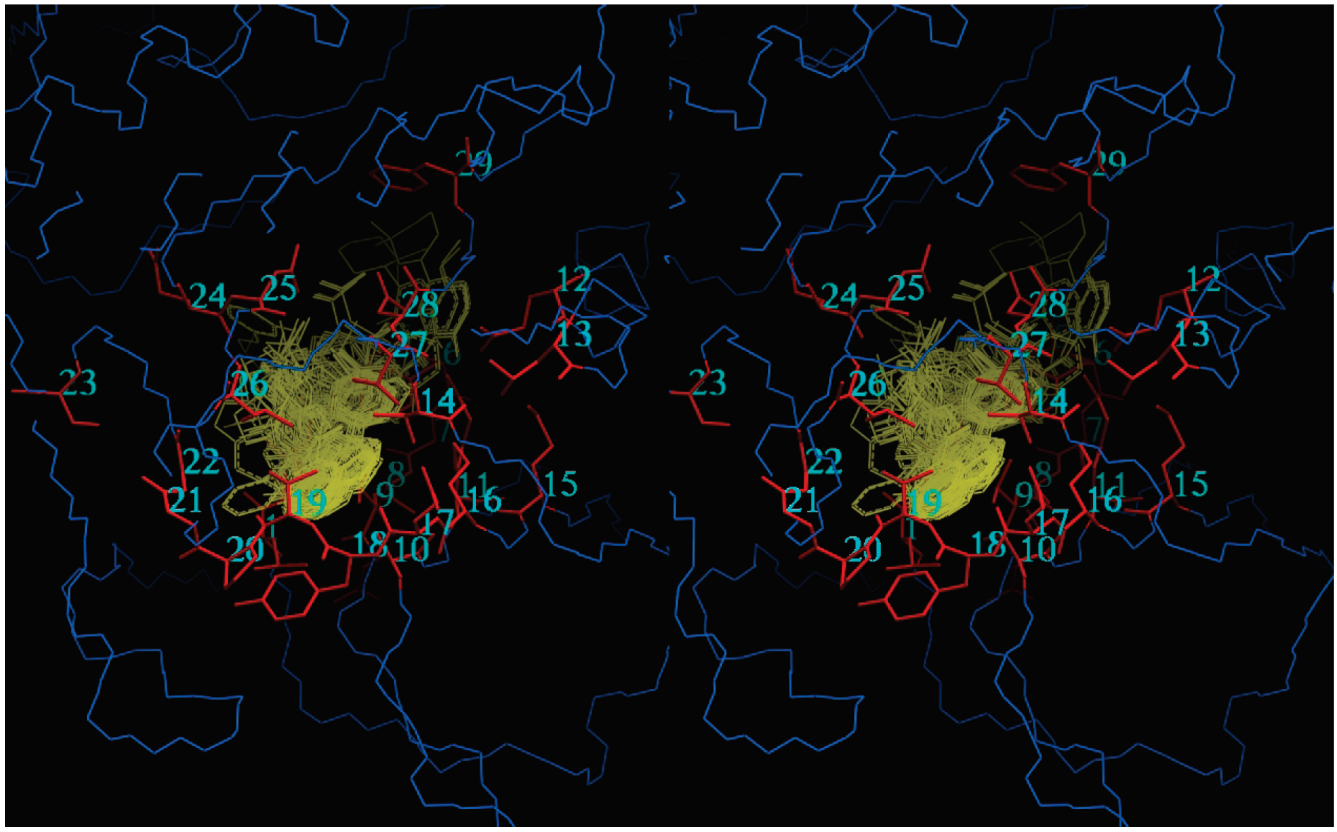

Figure 2. Placement of the 29 residues on the active site of human PKA (PDB entry 1STC). Ligands cocrystallized with PKA are shown in yellow wire to indicate the location of the binding site.

Table 3. Summary of the Data Sets Used in This Study

data set	endpoint	compounds	unique kinases in QSAR		kinase pairs in QSAR	cross-validated R2	important residues
Dundee	% Inh	200−900	50	mixed	1225	0.77	17,26,16,17,18,23
Karaman et al.	K_d	60	161	mixed	12880	0.85	16,21,22,17,5,6
Bamborough et al.	%control	577	181	mixed	16290	0.82	24,16,14,17
Federov et al.	Temperature shift.	157	49	Ser/Thr	1176	0.78	26,17,16,9,22
Melnick et al.	IC50	936	28	Tyr	378	0.81	16,18
Combined	N/A	N/A	262	mixed	24945	0.70	16,27,24,22,8,17,19

only two points. To avoid meaningless high similarities of this type, we included in the QSAR only those pairs of kinases where both kinases had at least 5 noninactive compounds. We did a similar filtering for Melnick et al. using $-\log(IC_{50}) = -1$ as the criterion of "inactive" and a similar filtering for Federov et al. using an absolute temperature shift of 2 Celsius as "inactive."

Henceforth we will refer to the Pearson correlation of the binding profile of two kinases as the "observed profile similarity". The distributions of the observed profile similarity for the data sets are shown in Figure 5. It is clear that they can be very different.

Combined Data Set. An anonymous reviewer suggested that we attempt a combined model. This requires constructing a data set called "Combined" that is the union of the data over the five data sets. Since the distribution of observed profile similarities of the data sets is so different, one must normalize each data set before combining them. The simplest way is to modify the original activities such that the mean of each data set is zero and the standard deviation is 1.0. Thus the unit of the combined data set would be in units of "zscore" rather than the Pearson correlation. If a pair of kinases is in more than one data set, we take the mean zscore. The data sets we have are mostly disjoint: only 32% of the union of 24945 unique kinase pairs occur in more than one data set.

QSAR Method. The point of our QSAR approach is to statistically relate the observed pairwise profile similarities to the descriptors generated from the pairwise comparison of alignments, thus generating a set of rules that could predict profile similarities for new pairs of kinases represented by descriptors. Random forest[38] is an ensemble recursive partitioning method that constructs predictions by averaging over multiple "trees". Each tree in the forest is constructed from a different bagged subset of the training set and at each branch point of the tree the method chooses from a random subset of the descriptors. Generally, we use 100 trees because using more does not improve the results. We used the R implementation of random forest (http://cran.r-project.org/src/contrib/Descriptions/randomForest.html). In our experience, random forest gives the best cross-validated predictions compared to other major QSAR methods and is the least sensitive to adjustable parameters. Recursive partitioning methods like random forest have the advantages that not all the cases have to be fit by one model, coupling between descriptors is naturally handled, and it is not assumed that

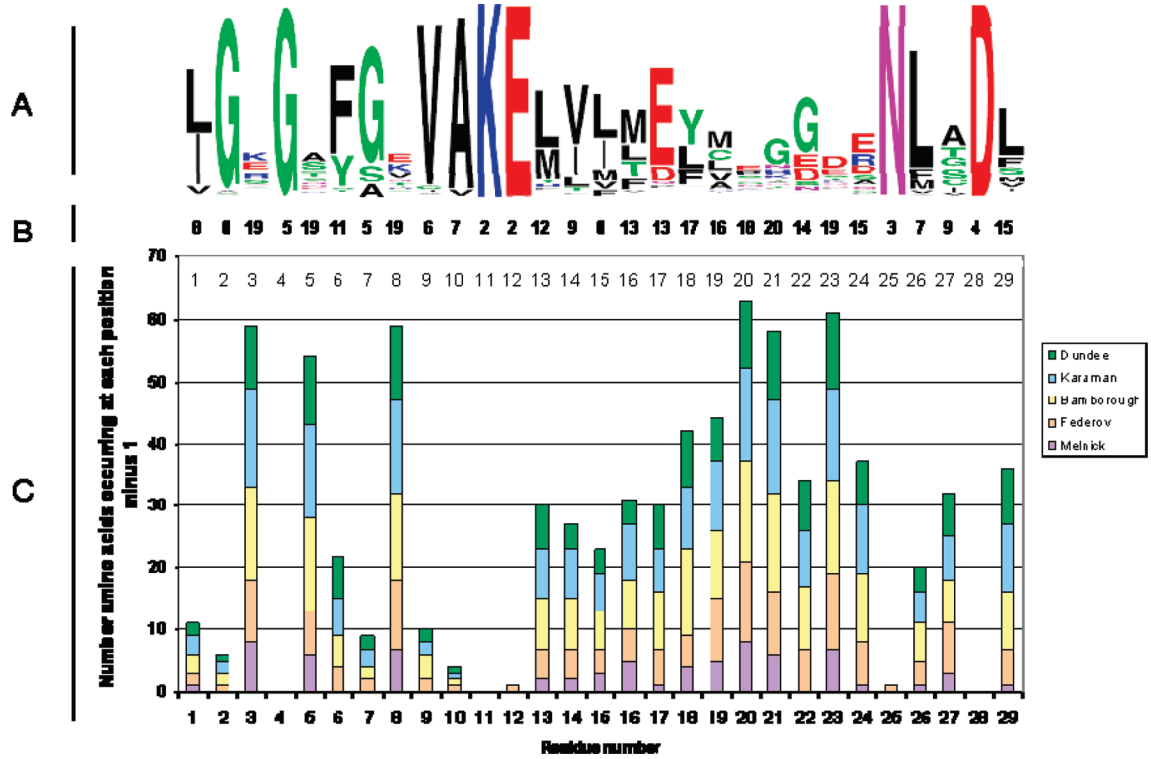

Figure 3. A: WebLogo plot (refs 39, 40) of amino acid frequency at each position for the 411 human kinases considered in the models. B: Number of unique amino acids occurring at each position for the 411 human kinases considered. C: Number of amino acids minus one occurring at each position for kinases included in each experimental data set.

the activity is a linear function of the descriptors. Also, recursive partitioning methods are not affected by having large numbers of irrelevant or correlated descriptors, so descriptor elimination is not necessary.

The importance of a descriptor for a random forest model may be gotten from the "out-of-bag" predictions during model building (on the average bagging leaves out about one-third of the cases). Each descriptor is in turn randomly reassigned to the wrong object (here a pair of kinases), and the accuracy of the prediction (over multiple trees) is monitored. The out-of-bag prediction accuracy will become much worse when an important descriptor is permuted but will change little when an unimportant one is.

One measure of the ability of a QSAR model to predict is by cross-validation. In this case, we randomly select half of the kinase pairs to form a "training set", generate a QSAR model, and then use the model to predict the remaining pairs, which form a "test set." Then we exchange the training and test sets. Repeating this 5 times gives us 5 predictions for each pair of kinases. The "cross-validated predicted profile similarity" for a kinase pair will be the mean over the 5 predictions. We can measure the overall goodness of the cross-validated predictions as the R^2 between predicted and observed profile similarities.

RESULTS

Observed Similarities versus Sequence Distance. Figure 6 shows the binding profile similarity as a function of the sequence distance of the 29 active site residues for Karaman et al. as an example. We would expect a good negative correlation between the binding profile similarity and the distance if all the (nonconserved) residues contributed equally to the binding profile. However, there is practically no overall correlation for any data set, consistent with the observation that kinases with similar sequences, even in the active site, can have very different binding profiles, and conversely that kinases with very different active sites overall can have similar binding profiles. This has been observed many times for the overall sequence homology of kinase domains. For example, Bamborough et al.[3] shows this on a large data set (for sequence similarity instead of sequence distance). This implies that there are probably only a few critical residues in the binding site that control the binding profile. The fact that very close sequences tend to have similar binding profiles, as noted by Bamborough et al., does not negate this idea; very similar sequences are likely to have critical residues that are identical.

Cross-Validation. The R^2 for the observed profile similarity and the mean of 5 leave-half-out cross-validated predicted profile similarities is shown in Table 3. An example is shown in Figure 7. Clearly, all individual models are internally self-consistent, the lowest R^2 for comparison of observed and cross-validated predicted profile similarity being 0.77. The Combined model is significantly less internally consistent than any individual model, with the R^2 being 0.70. Still, on an absolute basis, this is a reasonably large cross-validated R^2.

Descriptor Importances. Descriptor importances are in Table 4 based on QSAR models including all data for each data set. The scale of the descriptor importances depends on the size of the data set and the distribution of profile similarities, and so are valid only within a data set. By creating versions of each data set where the activities are randomly assigned to the wrong pairs, and looking at the maximum absolute value of descriptor importances from those version, we can get an idea of

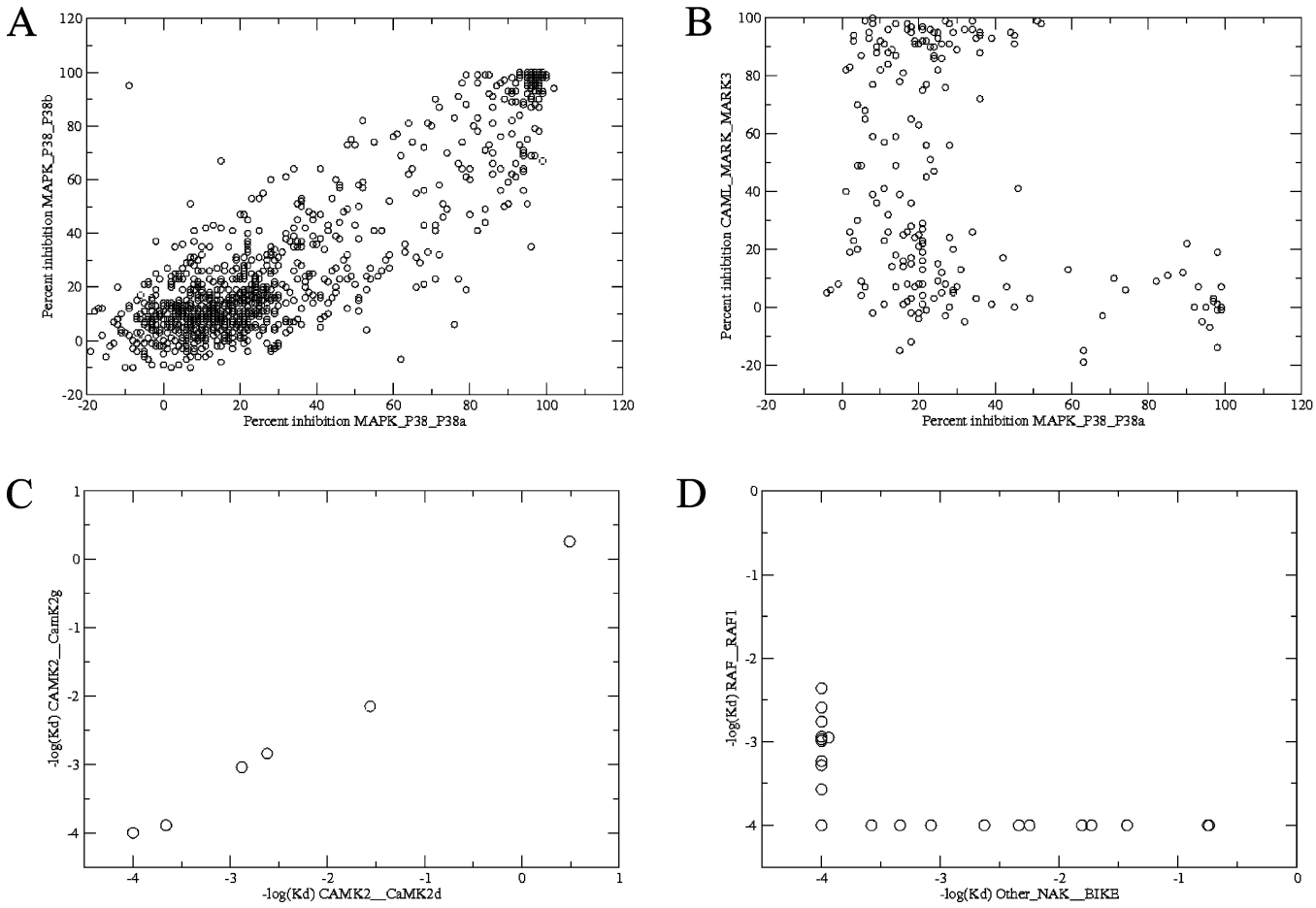

Figure 4. Extreme examples of correlations of percent inhibitions on kinase pairs. A and B from the Dundee set: MAPK_P38_P38a vs MAPK_P38_P38b (Pearson correlation 0.88) and CAMKL_MARK_MARK3 vs MAPK_P38_P38a (correlation −0.30). C and D from the Karaman et al. data set: CAMK2__CaMK2g vs CAMK2__CaMK2d (correlation 0.99) and RAF__RAF1 vs Other_NAK__BIKE (correlation −0.30). In this graph, each circle represents a compound.

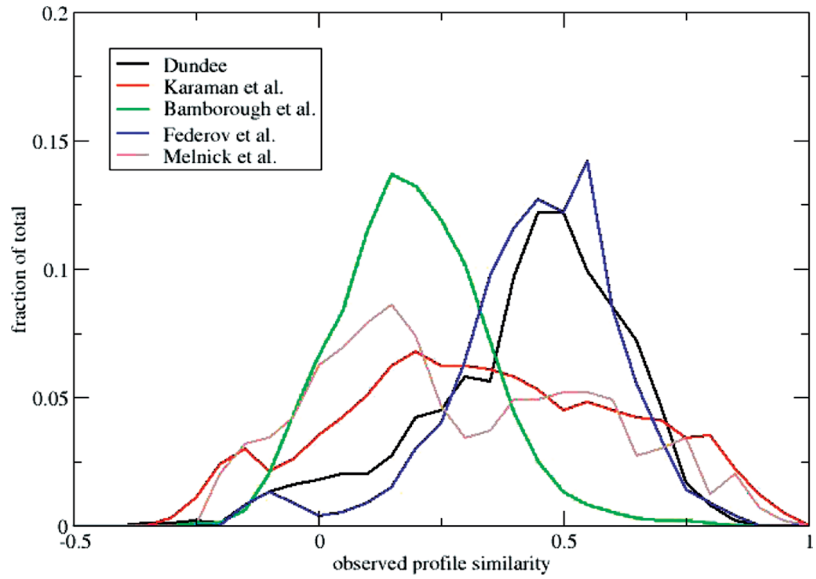

Figure 5. The distribution of observed profile similarities for the data sets.

when a descriptor importance is too small to be meaningful for that data set. We have also tested the stability of the descriptor importances to perturbations such as leaving out 20% of the kinase pairs. Only the descriptors with meaningful importances are shown in Table 4. Negative importances indicate that most descriptors are anticorrelated with the activity; this is as expected, the smaller the absolute differences in residue attributes between two kinases, the larger the similarity in binding profile between them is likely to be. The obvious exception is for the "IDENTITY" descriptor (e.g., in Bamborough et al., Federov et al. and Melnick et al.); if residues are identical in a pair of kinases they would be expected to have higher similarities in binding profile.

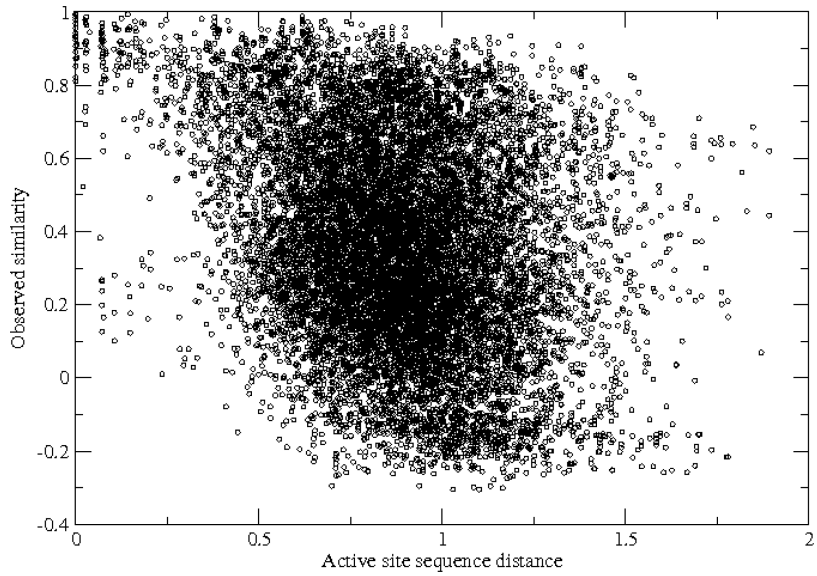

Figure 6. Observed pairwise profile similarities vs the sequence distance between the active sites of the kinases from Karaman et al. Each circle represents a pair of kinases. The active site distance was based on the 29 residues in Table 1 and was calculated using the PROTDIST module of PHYLIP (http://evolution.genetics.washington.edu/phylip.html).

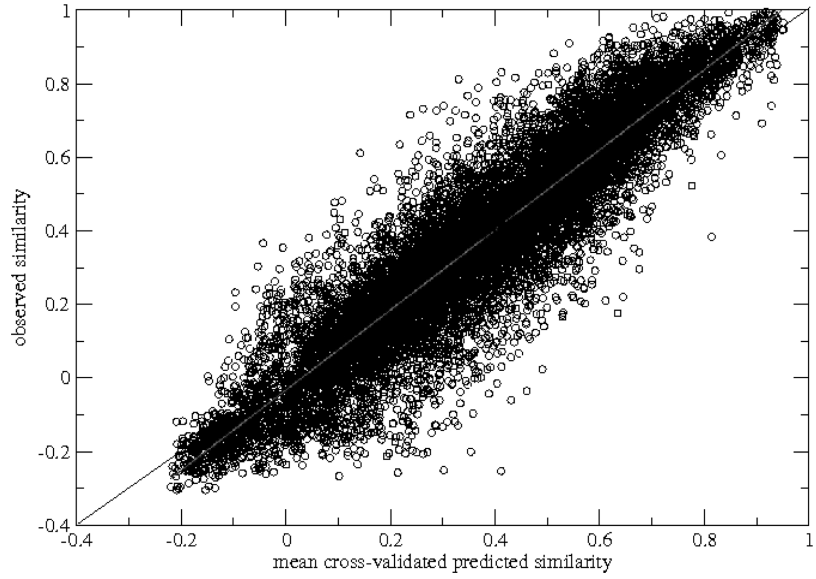

Figure 7. Observed profile similarity vs the cross-validated predicted similarity for Karaman et al. The predicted similarity for each pair represents an average over 5 trials where that pair was left out of the model. The black line is the diagonal and the red line is the best linear regression through the points. Each circle represents a pair of kinases. The R^2 for this scatterplot is 0.85.

Seven residues that almost never vary (2, 4, 10, 11, 12, 25, 28) would not be expected to be indicated as important in a QSAR study. As expected, only a few of the remaining 22 nonconserved residues in the active site have a large effect. All models include the gatekeeper residue, residue 16, as one important residue, and point to some flavor of polarity (e.g., DG_o_w, Polarity2, CONSENSUS_HYDROPHOBICITY) and its size (e.g., vdw_vol) as important variables. These variables distinguish among the residues that often occur in this position: Thr, Phe, Val, Met, or Leu. For Karaman et al. and Melnick et al., residue 16 is by far the most important residue. In the Dundee and Federov data sets, residue 16 is important but residues 17 and 26 appear more important. Also, in the Bamborough data sets, residues 24 and 14 appear important. We have not been able to discern an obvious reason why the models have the preferences they do: the type of experiment, the number of compounds, the substrate-specificity (Ser/Thr vs Tyr vs mixed), or the amino acid composition of key residues do not seem to be predictive. Not surprisingly, the combined model shows that residue 16, which is common among all individual models, is most important. Some of the important residues in the Combined model are in common with individual data sets (e.g., 17 and 24), but there is at least one new one (27).

Prediction of Profile Similarities for the Kinome. We created QSAR models using all the data in each data set and then used these models to predict the profile similarities for all kinase pairs in the kinome for which we have alignments. The predictions are in Supporting Information. The agreement of the predictions between models as measured by R^2 is shown in Table 5. Ideally, if the models have a good handle on what makes kinases bind compounds differently, all models would make similar predictions, but unfortunately the predictions do not agree to any significant

Table 4. Important Residue/Properties in Determining Binding Profile Similarity

Dundee	
17_DG_o_w	−0.670
23_HYDROPHOBICITY	−0.752
18_vdw_vol	−0.829
17_AROMATIC	−1.000
26_Polarity2	−1.051
16_vdw_vol	−1.302
26_MOL_WT	−1.345
26_DG_o_w	−2.127
26_vdw_vol	−2.773
17_FLEXIBILITY	−5.482
Karaman et al.	
6_Polarity	−10.137
5_pI	−11.199
22_HYDROPHOBICITY	−11.767
17_HYDROPHOBICITY	−12.815
16_FLEXIBILITY	−13.461
22_MUTABILITY	−14.342
21_MUTABILITY	−17.904
16_logD	−25.685
16_HYDROPHOBICITY	−33.438
16_CONSENSUS_HYDROPHOBICITY	−50.713
16_Polarity2	−118.067
16_DG_o_w	−135.195
Bamborough et al.	
24_IDENTITY	3.603
17_HYDROPHOBICITY	−3.343
16_DG_o_w	−3.481
24_pI	−3.630
14_Polarity	−3.799
24_vdw_vol	−4.136
16_Polarity2	−4.459
14_DG_o_w	−5.096
16_FLEXIBILITY	−5.144
27_Bulkiness	−5.694
16_MOL_WT	−6.288
14_CONSENSUS_HYDROPHOBICITY	−6.652
24_HYDROPHOBICITY	−7.220
16_logD	−8.486
24_CONSENSUS_HYDROPHOBICITY	−10.347
14_pI	−11.101
16_vdw_vol	−18.031
24_DG_o_w	−22.828
Federov et al.	
9_IDENTITY	1.551
22_pI	−0.449
26_MUTABILITY	−0.488
9_HYDROPHOBICITY	−0.497
9_MUTABILITY	−0.582
9_Polarity2	−0.629
26_Bulkiness	−0.720
26_DG_o_w	−0.910
26_logD	−0.982
26_vdw_vol	−1.230
16_vdw_vol	−1.330
17_BRANCHED	−2.545
26_Polarity2	−3.828
Melnick et al.	
16_IDENTITY	1.344
18_MUTABILITY	−0.542
16_CONSENSUS_HYDROPHOBICITY	−0.542
18_DG_o_w	−0.548
16_HYDROPHOBICITY	−0.825
16_MOL_WT	−0.933
16_DG_o_w	−0.999
16_FLEXIBILITY	−1.629
16_pI	−1.946
16_MUTABILITY	−4.346
Combined	
14_Polarity2	−197.936
19_FLEXIBILITY	−198.257
17_HYDROPHOBICITY	−235.730
8_DG_o_w	−236.512
22_MUTABILITY	−250.478
24_DG_o_w	−274.009
16_HYDROPHOBICITY	−280.858
16_CONSENSUS_HYDROPHOBICITY	−306.159
16_DG_o_w	−369.989
27_Bulkiness	−508.531
16_logD	−529.152
16_vdw_vol	−818.432
16_Polarity2	−1006.307

extent among the individual models. Figure 8 compares the predictions of kinome pairs that are outside the training set for one pair of models as an example. The Combined model has reasonably large R^2 with the Karaman et al. and Bamborough et al. models. This is not surprising because most of the data in the Combined data set comes from those large data sets.

Given that the data sets have different subsets of kinases, it might not be surprising that the models might give different predictions, but we believe the better part of the difference is more fundamental than that and has to do with the lack of agreement between individual data sets on the same pairs of kinases. Table 6 shows the R^2 for all pairs of data sets. For example, the Karaman et al. set and the Dundee set contain 190 kinase pairs in common. The comparison of their observed profile similarities is shown in Figure 9. While there is an overall trend, where one can distinguish very dissimilar kinases from others, the overall R^2 for this comparison is only 0.4, and that is one of the larger R^2 values. The large number of off-diagonal pairs in such plots would indicate that this is a probably a real issue and not just an artifact of, say, converting a few kinases to the wrong Manning et al. nomenclature. The use of rank correlations does not make the R^2 significantly more favorable, indicating that the difference in distribution of profile similarities or nonlinear mapping of observations from one data set to another does not account for the poor R^2. Given the disagreements between data sets, it is to be expected that the cross-validated R^2 for the Combined model is lower than any individual model. We see in Table 6 that the combined data set does have a reasonably large R^2 with each individual data set, but this is a reflection of the fact that most kinase pairs are found in only one data set. For example, most of the kinase pairs in the Combined data set that are common with, say, Dundee, are just zscore-rescaled versions of Dundee pairs.

DISCUSSION

We have created a QSAR methodology to predict the similarity of binding profiles in kinases and used this methodology to build QSAR models using five data sets. This methodology is in contrast to exercises where QSAR models of compounds on individual kinases are built,[17,28,32] or the binding properties of individual kinase active sites are compared.[29–31] One should be aware of the limitations and assumptions of our approach:

1. The models can predict only how similar two kinases are likely to be in terms of their overall binding profile but are mute about the binding of a particular compound on a particular kinase.
2. The Pearson correlation measure of profile similarity ignores absolute value of binding, and takes into account only the binding of compounds relative to each other. Thus if the percent inhibitions of kinase B are shifted relative to kinase A by being 50% lower, the Pearson correlation of kinase A and B would be 1.0, despite kinase B binding the compounds more weakly. Hence if A and B are "similar" by the Pearson correlation that does not necessarily mean they bind the same compounds tightly. This is in contrast to another type of similarity, the "Tanimoto similarity" used by Bamborough et al.,[3] where a certain level of potency is set to call a molecule "active" on a particular kinase, and the Tanimoto similarity

Table 5. R^2 for Predicted Profile Similarity of Kinase Pairs for the Kinome Compared between Data Sets, Not Counting Kinase Pairs in Either Training Set

	Dundee	Karaman et al.	Bamborough et al.	Federov et al.	Melnick et al.	Combined
Dundee		0.07	0.08	0.19	0.02	0.14
Karaman et al.			0.27	0.22	0.10	0.53
Bamborough et al.				0.20	0.12	0.56
Federov et al.					0.08	0.21
Melnick et al.						0.10
Combined						

measures what fraction of active compounds two kinases have in common. Our experience comparing Pearson and Tanimoto similarities on our data sets is that the Pearson gives a higher cross-validated R^2, making it more suitable for QSAR. Also, the Pearson correlation does not require introducing an arbitrary cutoff.

3. We assume that all the important forces for determining a profile are the result of the residues immediately surrounding the ATP-binding site. This seems reasonable and information about the active site produces an internally consistent model. Further, the difficulty in producing a sequence alignment for entire kinase sequences precludes us from investigating the effect of more residues.
4. As with any QSAR model, one must have a good sample of the domain of interest one wants to model. Getting data sets with enough compounds tested on enough representative kinases can be an issue. Also all a QSAR model can do is summarize known information in a useable form. It does not necessarily give one mechanistic insight.

The models we generate appear internally self-consistent as measured by the cross-validated R^2, so there is no insurmountable problem with building models. This would imply that errors that might interfere with the model building, such as misassignment of kinases names to sequences, are probably minor. All models say that only a few of the residues in the active site control the relative binding of compounds, and agree that the gatekeeper is among them, although the models differ in which residues are most important. The gatekeeper residue has been shown to play a critical role in the selectivity of tyrosine kinase inhibitors. The size of this residue, which corresponds to residue 16 in our models, determines the size of the distal hydrophobic pocket. Most kinases feature a large hydrophobic residue in this position, such as Phe, Met, or Leu, and selectivity for kinases with a smaller amino acid, typically Val or Thr, can be obtained by placing large substituents in this region of the binding site; this feature is exploited by Gleevec (imatinib) and Iressa (gefitinib), among others, to achieve

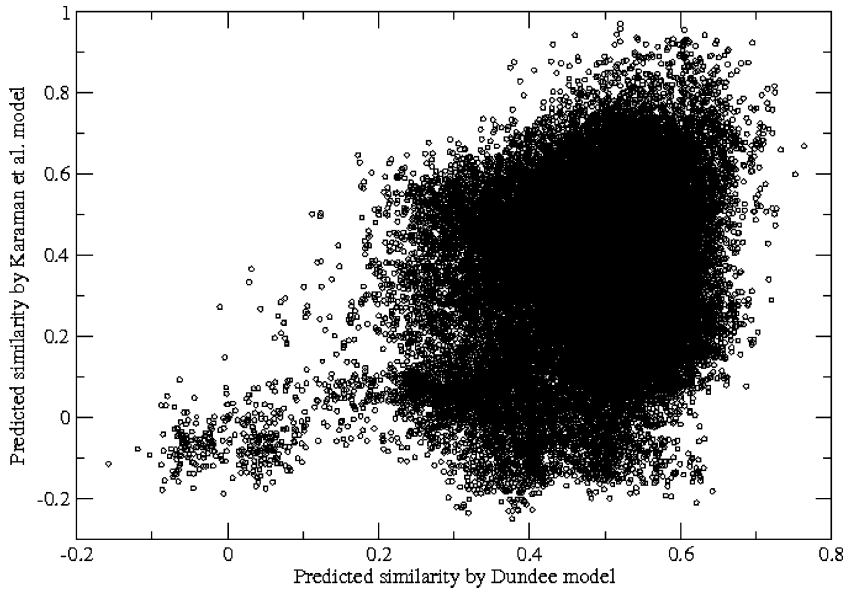

Figure 8. Comparison of predictions of kinome pairs by Karaman et al. and Dundee. We show only those pairs that are outside the training set of both models. Each circle represents a pair of kinases.

Table 6. R^2 for Observed Activities of Kinase Pairs Compared between Data Sets[a]

	Dundee	Karaman et al.	Bamborough et al.	Federov et al.	Melnick et al.	Combined
Dundee		0.40 (190)	0.25 (406)	0.75 (28)	(0)	0.81 (1225)
Karaman et al.			0.28 (7620)	0.63 (231)	0.16 (210)	0.70 (12880)
Bamborough et al.				0.31 (528)	0.19 (325)	0.73 (16290)
Federov et al.					(0)	0.77 (1176)
Melnick et al.						0.46 (378)
Combined						

[a] The number in parentheses is the number of kinase pairs the data sets have in common.

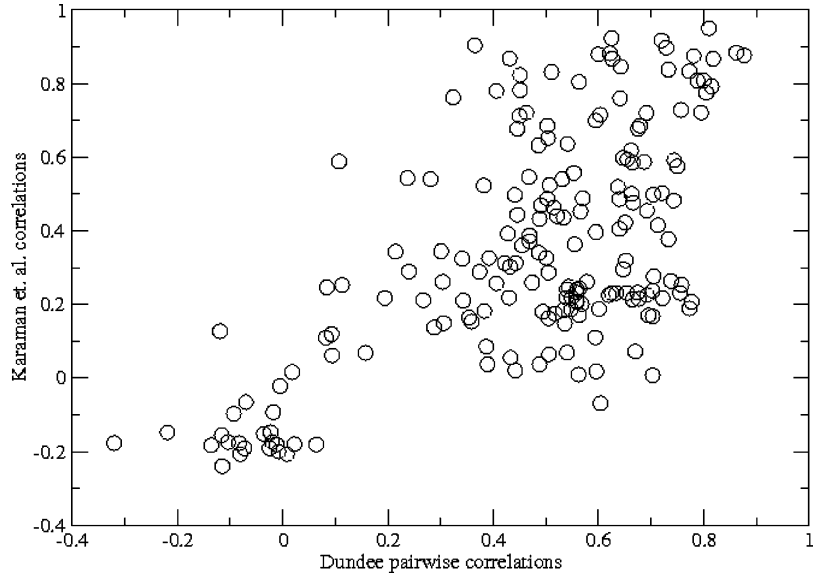

Figure 9. Observed pairwise similarities in Karaman et al. vs Dundee. Each circle represents a pair of kinases.

selectivity. The fact that the effect of the gatekeeper is consistently seen is encouraging.

It is not surprising that QSAR models should differ when the nature of the input data is different. As far as we know, no one has made this type of comparison between experimental data sets, and we were surprised how disjoint the data sets are and, where the sets have common kinases, how little any two experimental data sets agree with each other about which kinases are similar to which others. Certainly, the same issue will affect any QSAR study of kinases.

The reason that data sets disagree could be a number of things: differences in the endpoint of the assay, (e.g., whether binding or inhibition is being measured), the experimental conditions (e.g., the ATP concentration), the rather imprecise nature of percent inhibitions versus IC_{50} values, whether the compounds tested explore different parts of the active site, etc. Picking among these possibilities is purely speculative at this point since the number of data sets is very small, and in some cases, the nature of the molecules is kept proprietary. In retrospect, perhaps the fact that different laboratories run assays for different purposes makes the differing results not so surprising.

Given that the models disagree, we cannot say that we have a definitive model that will cover all types of experimental conditions, and it may be necessary to generate a model for a specific type of experiment and confine predictions to that type of experiment. Our attempt to generate a Combined model by taking the union of all the data sets was successful in the sense that such a model contains many more kinases than any individual model and is reasonably self-consistent by cross-validation, although significantly less self-consistent than any individual model. The fact that it is necessary to express the activities in the Combined model as a zscore, because the individual data sets have very different distributions in original Pearson correlation units, is a slight impediment, in the sense that we cannot directly return a prediction as an absolute Pearson similarity. However, the Combined model still can order kinase pairs by profile similarity, which means it could potentially be useful, for instance, for selecting similar kinases for a counter screen. If necessary one could back-transform the zscore units to a reasonable range of Pearson units to get an absolute estimate of whether two kinases were "similar." Whether the combined model can be considered usefully "universal" is debatable for two reasons. First, because the Combined data set is dominated by the two larger individual data sets Karaman et al. and Bamborough et al. Second, because of the large multiple differences in the original experiments, as noted, above it is not clear that combining data from the separate data sets is meaningful.

Obviously for a QSAR model of the type we describe here one would like a complete data matrix of precise data (e.g., K_d or IC_{50}) on many diverse compounds tested on many diverse kinases *under the same experimental conditions*, but currently, we do not have access to data that meets this ideal. There are now consortia that will, for a fee, generate IC_{50} values for a large number of compounds on a large number of kinases, so getting IC_{50} data is in the realm of possibility for further model development.

ACKNOWLEDGMENT

We thank the investigators who provided experimental data sets in machine-readable form. We also gratefully acknowledge the resources of the Division of Signal Transduction Therapy (DSTT), University of Dundee, Sir Philip Cohen FRS FRSE, Co-director.

Supporting Information Available: Alignment of active site residues for 411 kinases, list of amino acid properties used for QSAR descriptors, and observed profile similarities for all kinase pairs in 6 data sets, predicted profile similarities for all kinase pairs for 6 models, plus the active site sequence distances. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human. *Science* **2002**, *298*, 1912–1934.

(2) Bain, J.; Cummings, L.; Elliot, M; Shpiro, N.; Hastie, J.; McLauchian, H.; Klevernic, I.; Arthur, S.; Alessi, D.; Cohen, P. The selectivity of

protein kinase inhibitors; a further update. *Biochem. J.* **2007**, *48*, 297–315.
(3) Bamborough, P.; Drewry, D.; Harper, G.; Smith, G. K.; Schneider, K. Assessment of chemical coverage of kinome space and its implications for kinase drug discovery. *J. Med. Chem.* **2008**, *51*, 7998–7914.
(4) Federov, O.; Marsden, B.; Pogacic, V.; Rellos, P.; Muller, S.; Bullock, A. N.; Schwaller, J.; Sundstrom, M.; Knapp, S. A systematic interaction map of validated kinase inhibitors with Ser/Thr kinases. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 20523–20528.
(5) Karaman, M. W.; Herrgard, S.; Treiber, D. K.; Gallant, P.; Atteridge, C. E.; Campbell, B. T.; Chan, K. W.; Cicerci, P.; Davis, M. I.; Edeen, P. T.; Faraoni, R.; Floyd, M.; Hunt, J. P.; Lockhart, D. J.; Milanov, Z. V.; Morrison, M. J.; Pallares, G.; Patel, H. K.; Pritchard, S.; Wodicka, L. M.; Zarrinkar, P. P. A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **2008**, *26*, 127–132.
(6) Melnick, J. S.; Janes, J.; Kim, S.; Chang, J. Y.; Sipes, D. G.; Gunderson, D.; Jarnes, L.; Matzen, J. T.; Garcia, M. E.; Hood, T. L.; Beigi, R.; Xia, G.; Harig, R. A.; Asatryan, H.; Yan, S. F.; Zhou, Y.; Gu, X.-J.; Saadat, A.; Zhou, V.; King, F. J.; Shaw, C. M.; Su, A. I.; Downs, R.; Gray, N. S.; Schultz, P. G.; Warmuth, M.; Caldwell, J. S. An efficient rapid system for profiling the cellular activities of molecular libraries. *Proc. Natl. Acad. U. S. A.* **2006**, *103*, 3153–3158.
(7) Bonn, S.; Herrero, S; Breitenlechner, C. B.; Erlbruch, A.; Lehmann, W.; Engh, R. A.; Gassel, M.; Bossemeyer, D. Structural analysis of protein kinase A mutants with rho-kinase inhibitor specificity. *Biol. Chem.* **2006**, *281*, 24818–24830.
(8) Cheney, I. W.; Yan, S.; Appleby, T.; Walker, H.; Vo, T.; Yao, N.; Hamatake, R.; Hong, Z.; Wu, J. Z. Identification and structure-activity relationships of substituted pyridones as inhibitors of Pim-1 kinase. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 1679–1683.
(9) Cherry, M.; Williams, D. H. Recent kinase and kinase inhibitor X-ray structures: Mechanisms of inhibition and selectivity insights. *Curr. Med. Chem.* **2004**, *11*, 663–673.
(10) Cohen, M. S.; Zhang, C.; Shokat, K. M.; Taunton, J. Structural bioinformatics-based design of selective, irreversible kinase inhibitors. *Science* **2005**, *308*, 1318–1321.
(11) Debreczeni, J. E.; Bullock, A. N.; Atilla, G. E.; Williams, D. S.; Bregman, H.; Knapp, S.; Meggers, E. Ruthenium half-sandwich complexes bound to protein kinase Pim-1. *Angew. Chem., Int. Ed.* **2006**, *45*, 1580–1585.
(12) Emrick, M. A.; Lee, T.; Starkey, P. J.; Mumby, M. C.; Resing, K. A.; Ahn, N. G. The gatekeeper residue controls autoactivation of ERK2 via a pathway of intramolecular connectivity. *Proc. Natl. Acad. Sci U. S. A* **2006**, *103*, 18101–18106.
(13) Fitzgerald, C. E.; Patel, S. B.; Becker, J. W.; Cameron, P. M.; Zaller, D.; Pikounis, V. B.; O'Keefe, S. J.; Scapin, G. Structural basis for p38α MAP kinase quinazolinone and pyridol-pyrimidine inhibitor specificity. *Nat. Struct. Biol.* **2003**, *10*, 764–769.
(14) Goldberg, D. R.; Hao, M.-H.; Qian, K. C.; Swinamer, A. D.; Gao, D. A.; Xiong, Z.; Sarko, C.; Berry, A.; Lord, J.; Magolda, R. L.; Fadra, T.; Kroe, R. R.; Kukula, A.; Madwed, J. B.; Martin, L.; Pargellis, C.; Skow, D.; Song, J. J.; Tan, Z.; Torcelli, C.A.; Zimmitti, C. S.; Yee, N. K.; Moss, N. Discovery and optimization of p38 inhibitors via computer-assisted drug design. *J. Med. Chem.* **2007**, *50*, 4016–4026.
(15) Hamdouchi, C.; Zhong, B.; Mendoza, J.; Collins, E.; Jaramillo, C.; De Diego, J. E.; Robertson, D.; Spencer, C. D.; Anderson, B. D.; Watkins, S. A.; Zhang, F.; Brooks, H. B. Structure-based design of a new class of highly selective aminoimiazo[1,2-*a*]pyridine-based inhibitors of cyclin dependent kinases. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1943–1947.
(16) Kothe, M.; Kohls, D.; Low, S.; Coli, R.; Rennie, G. R.; Feru, F.; Kuhn, C.; Ding, Y.-H. Selectivity-determining residues in Plk1. *Chem. Biol. Drug Des.* **2007**, *70*, 540–546.
(17) Myrianthopoulos, V.; Magiatis, P.; Ferandin, Y.; Skalt Sounis, A. J.; Meijer, L.; Mikros, E. A integrated computational approach to the phenomenon of potent and selective inhibition of Aurora kinases B and C by a series of 7-substituted indirubins. *J. Med. Chem.* **2007**, *50*, 4027–4037.
(18) Pande, V.; Ramos, M. J. Structural basis for the GSK-3b binding affinity and selectivity against CDK-2 of 1-(4-aminofurazan-3yl)-5-dialkylaminomethyl-1-*H*-[1,2,3] triazole-4-carboxylic acid derivatives. *Biorg. Med. Chem. Lett.* **2005**, *15*, 5129–5135.
(19) Azam, M.; Seeliger, M. A.; Gray, N. S.; Kuriyan, J.; Daley, G. Q. Activation of tyrosine kinases by mutation of the gatekeeper threonine. *Nat. Struct. Mol. Biol.* **2008**, *15*, 1109–1118.
(20) Aronov, A. M.; Murcko, M. A. Toward a pharmacophore for kinase frequent hitters. *J. Med. Chem.* **2004**, *47*, 5616–5619.
(21) Aronov, A. M.; McClain, B.; Moody, C. S.; Murcko, M. A. Kinase-likeness and kinase-privileged fragments: Toward a virtual polypharmacology. *J. Med. Chem.* **2008**, *51*, 1214–1222.
(22) Sprous, D. G.; Zhang, J.; Zhang, L.; Wang, Z.; Tepper, M. A. Kinase inhibitor recognition by use of a multivariable QSAR model. *J. Mol. Graph. Model.* **2006**, *24*, 278–295.
(23) Sutherland, J. J.; Higgs, R. E.; Watson, I.; Vieth, M. Chemical fragments as foundations for understanding target space and activity prediction. *J. Med. Chem.* **2008**, *51*, 2689–2700.
(24) Gozalbes, R.; Simon, L.; Froloff, N.; Sartori, E.; Monteils, C.; Baudelle, R. Development and experimental validation of a docking strategy for the generation of kinase-targeted libraries. *J. Med. Chem.* **2008**, *51*, 3124–3132.
(25) Muegge, I.; Enyedy, I. J. Virtual screening for kinase targets. *Curr. Med. Chem.* **2004**, *11*, 693–707.
(26) Rockey, W. M.; Elcock, A. H. Rapid computational identification of the targets of protein kinase inhibitors. *J. Med. Chem.* **2005**, *48*, 4138–4152.
(27) Zahler, S.; Tietze, S.; Totzke, F.; Kubbutat, M.; Meijer, L; Vollmar, A. M.; Apostolakis, J. Inverse in silico screening for identification of kinase inhibitor targets. *Chem. Biol.* **2007**, *14*, 1207–1214.
(28) Heady, L.; Fernandez-Serra, M.; Mancera, R. L.; Joyce, S.; Venkitaraman, A. R.; Artacho, E.; Skylarkis, C.-K.; Ciacchi, L. C.; Payne, M. C. Novel structural features of CDK inhibition revealed by ab initio computational method combined with dynamic simulations. *J. Med. Chem.* **2006**, *49*, 5141–5153.
(29) Sheinerman, F. B.; Giraud, E.; Laoui, A. High affinity targets of protein kinase inhibitors have similar residues at the positions energetically important for binding. *J. Mol. Biol.* **2005**, *352*, 1134–1156.
(30) Chen, J.; Zhang, X.; Fernandez, A. Molecular basis for specificity in the druggable kinome: Sequence-based analysis. *Bioinformatics* **2007**, *23*, 563–572.
(31) Kinnings, S. L.; Jackson, R. M. Binding site similarity analysis for the functional classification of the protein kinase family. *J. Chem. Inf. Model.* **2009**, *49*, 318–329.
(32) Sciabola, S.; Stanton, R. V.; Wittkopp, S.; Wildman, S.; Moshinsky, D.; Potluri, S; Xi, H. Predicting kinase selectivity profiles using Free-Wilson QSAR analysis. *J. Chem. Inf. Model.* **2008**, *48*, 1851–1867.
(33) Graczyk, P. P. A Gini coefficient: a new way to express selectivity of kinase inhibitors against a family of kinases. *J. Med. Chem.* **2007**, *50*, 5773–5779.
(34) Fabian, M. A.; Biggs, W. H., III; Trieber, 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.
(35) Vieth, M.; Sutherland, J. J.; Robertson, D. H.; Cambell, R. M. Kinomics: Characterizing the therapeutically validated kinase space. *Drug Discovery Today* **2005**, *10*, 839–846.
(36) RCSB Protein Data Bank. www.rscb.org. (accessed July 2009).
(37) Abagayan, R. A.; Totrov, M. M.; Kuznetsov, D. N. ICM—A new method for protein modeling and design. Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488–506.
(38) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
(39) Schneider, T. D.; Stephens, R. M. Sequence Logos: A new way to display consensus sequences. *Nucleic Acids Res.* **1990**, *18*, 6097–6100.
(40) Crooks, G. E.; Hon, G.; Chandonia, J.-M.; Brenner, S. E. WebLogo: A sequence logo generator. *Genome Res.* **2004**, *14*, 1188–1190.

CI900176Y