Classifying Proteins by Common Conserved Motifs to Control Ligand Binding Specificity

Peter Anderson, Kevin Turner, Sidney Elmer, Vincent De Sapio, Joe Schoeniger, Diana Roe

Sandia National Laboratories, P.O. Box 969, MS 9291, Livermore, CA 94551

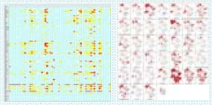
SAND2011-1940C

Abstract

A major challenge in drug design is controlling binding specificity. High specificity for drugs that target human proteins is often desirable to limit off-target binding. However, lower specificity may be desirable in certain cases, for example, for drugs that target a range of pathogen strains or mutations. We have developed the Common Conserved Motifs (CCM)[1,2] method to identify structural features of proteins and ligands that determine selectivity across a range of targets. In this study we used the data set from Karaman *et al*[3] for kinase/ligand binding data. We cluster a large set of human kinases and kinase inhibitors by their experimental binding profiles and identify *in silico* sets of cluster-specific features that influence narrow and broad inhibitor binding. Knowledge of these features has allowed us to predict and experimentally confirm specific kinase inhibition for ligands outside the training set. Our method has applications in both lead drug design and target selection.

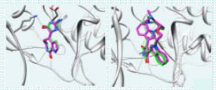
Methodology

1. Obtain tables of binding data for ligand binding to protein families



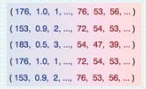
2. Cluster binding data based on Euclidian distances between vectors

3. Perform *in silico* docking of all ligands to all proteins



4. Identify intermolecular interaction motifs that are unique to each protein cluster

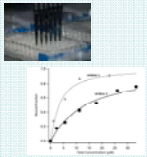
5. Represent distances between atoms of docked ligands and the unique cluster interaction motifs as interaction vectors



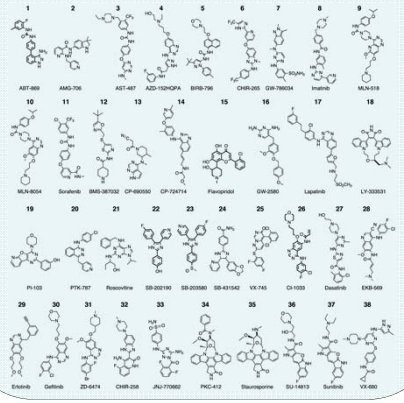
6. Construct a prediction model to associate vectors with experimental $\log(K_d)$ values or cluster numbers

7. Predict binding ability or cluster assignments for additional ligands

8. Experimentally verify binding affinity predictions

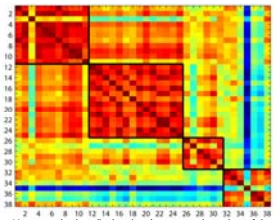


Ligand Data Set



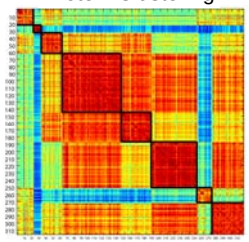
Example classification scheme for the Karaman *et al* set of 317 kinases and 38 ligands.[3] We clustered both the proteins with respect to ligand binding profiles (top left) and ligands by protein binding profiles (bottom left) using the Matlab software package[4]. The right panel shows the binding matrix for ligands vs. proteins ordered by the ligand and protein clusterings. A *k*-centers algorithm was used to cluster the kinases based on binding similarities. The Euclidean distance matrix associated with $\log(K_d)$ values of 38 ligands (i.e. 38 dimensions) was reduced to 3 principal components using PCA for the clustering metric, and the total number of clusters to generate was specified. Similarly the ligands were reduced to 3 principal components and clustered. Ordered heat maps were plotted based on different clusterings. Kinases (or ligands) of the same cluster are located contiguously in the grid. The results, particularly for 8 (shown) and 16 clusters, indicate that kinases within a cluster tend to share strong similarity in the binding metric relative to each other and weaker similarities in the binding metric relative to kinases in other clusters. Ligand clusters of 4 (shown) and 8 were robust.

Ligand Clustering



Heatmap (ordered) showing *k*-centers clustering of the ligands based on binding similarities for 4 clusters.

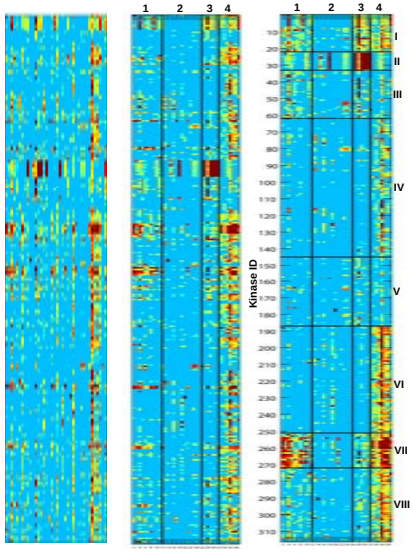
Protein Clustering



Kinase ID

Heatmap (ordered) showing *k*-centers clustering the kinases based on binding similarities for 8 clusters.

Binding Matrix Ordered by Clusters

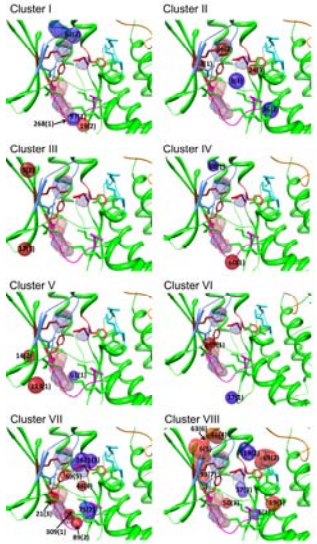


Unordered Binding Matrix

Binding matrix ordered by 4 ligand clusters

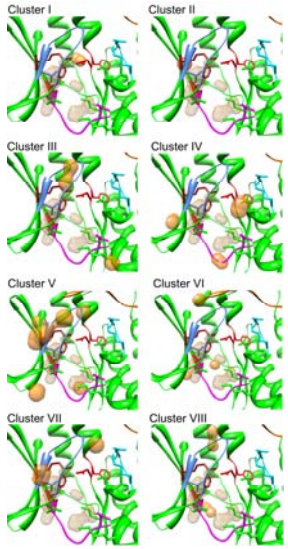
Binding matrix ordered by 4 ligand clusters and 8 protein clusters

Unique, Cluster-Specific Hydrogen Bonding Regions



Hydrogen bond donors
Hydrogen bond acceptors

Unique, Cluster-Specific Hydrophobic Interaction Regions



Predicting Kinase Inhibitors and Non-Inhibitors Using a Random Forest Model with Identified Unique Features as Descriptors

For each of the 3839 kinase-ligand complexes whose structures we predicted by docking, a 578-element vector was constructed. The first 77 elements consist of the distances between ligand hydrogen-bonding atoms and all 77 unique hydrogen bonding regions that we have identified. The next 400 elements are the distances between ligand atoms that form polar contacts with the kinase and the 400 identified unique kinase polar-interaction regions. The final 101 vector elements comprise the distances between the ligand atoms involved in hydrophobic interactions with the kinase and the 101 identified unique hydrophobic interaction regions. Each vector was assigned an **inhibitor** or **non-inhibitor** class label based on whether the experimental K_d for the complex is stronger than or weaker than 10 μM, respectively. The data set includes 892 inhibitors and 2947 non-inhibitors.

Following vector construction, we trained a random forest (RF) model using the R implementation of random forest. RF is an ensemble partitioning method that builds predictions by averaging over multiple decision trees. The 578-element vectors were treated as independent variables, and the corresponding **inhibitor** and **non-inhibitor** class labels were treated as dependent variables. The RF model ensemble used 500 trees and m_{try}=24.

Our RF model has an out-of-bag prediction accuracy of 76% for inhibitors and 83% for non-inhibitors.

The RF model was subsequently used to predict the inhibition ability of 9 compounds from outside the Karaman *et al* set: aloisine, NU-6102, SC-221409, SU-11274, D4426, quinoxaline1, tyrphostinA23, scytonemin, and dimethyladenine. These compounds were docked to CDK2, ZAP70, and PYK2, and the standard 578-element descriptor vectors were constructed. The vectors were fed into the RF model to obtain predictions of inhibition.

	Aloisine	NU-6102	SC-221409	SU-11274	D4426	Quinoxaline1	Tyrp.A23	Scytonemin	Dimethyl-adenine
CDK2	+	+	+	+	+	+	+	+	+
ZAP70	+	+	+	+	+	+	–	+	+
PYK2	+	–	–	–	+	+	–	+	–

+ predicted inhibitor **–** predicted non-inhibitor

Experimental Validation of Inhibitor Predictions

We employed electrospray ionization mass spectrometry (ESI-MS) to perform inhibition assays for each of the 9 compounds from outside our training set against kinases CDK2, ZAP70, and PYK2. These kinases are located in protein clusters IV, IV, and VIII, respectively. Each enzyme-inhibitor pair was analyzed at three inhibitor concentrations (0.1x, 1x, and 10x enzyme concentration) by calculating the reduction of the phosphorylation rate of substrate peptides relative to control reactions that took place under V_{max} conditions in the absence of inhibitor.

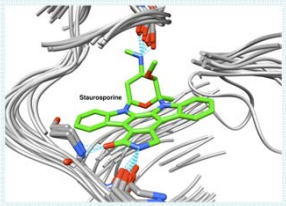
Experimentally measured ligand K_i values (nM)

Ligand	CDK2	ZAP70	PYK2
Aloisine	480	1	1820
NU-6102	810	1	2450
SC-221409	1200	110	2820
SU-11274	10	470	1780
D4426	360	10	980
Quinoxaline1	30	1910	720
TyrphostinA23	100	540	1350
Scytonemin	850	350	1260
Dimethyladenine	1580	1050	9330

What is the key feature of a broadly binding kinase inhibitor?

Hydrogen-bonding interactions between staurosporine and the kinase proteins only occur between the ligand and the protein backbone.

Key conclusion: broad binding may be facilitated by designing ligands to interact strongly with the invariant protein backbone, rather than the variable sidechains. Conversely, specific binding may be attributed to highly specific contacts between the ligand and the protein sidechains.



Broad binder (Staurosporine) binding to representatives from all 8 protein clusters

References

[1] Zhou, C.E.; Zemla, A.T.; Schoeniger, J.S.; Roe, D.C., Methods And Systems Of Common Motif And Countermeasure Discovery, *Patent Pending*
[2] Ecale Zhou, C.L.; Zemla, A.T.; Roe, D.; Young, M.; Lam, M.; Schoeniger, J.S.; Balhorn, R. (2005) Computational approaches for identification of conserved/unique binding pockets in the A chain of ricin, *Bioinformatics* **21**, 3089-3096.
[3] Karaman, M.W. *et al* (2008) A quantitative analysis of kinase inhibitor selectivity, *Nat. Biotechnol*, **26**, 127-132.
[4] Matlab: The MathWorks Natick, Massachusetts
[5]. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis, *J Comput. Chem.* **25**, 1605-1612.
[6] Autodock 4. , Scripps Research Institute, La Jolla, CA

Acknowledgements

This work was supported by the DTRA CB Basic Research Program and by the Laboratory Directed Research and Development program at Sandia National Laboratories.