SAND2010-5270P

Studying high efficacy mechanisms involved in controlling specificity in molecular recognition BRCALL08-L-2-0006

Diana Roe (PI) Joe Schoeniger (co-PI)
Principal Member of the Technical Staff
Sandia National Labs

DTRA Chemical and Biological Defense
Basic Research Technical Review

August 2010

Project Objective

- *Provide an improved fundamental understanding of the physical features (flexibility, shape and charge motifs) of proteins and ligands that determine their binding specificity; demonstrate that this understanding can be used to predict and control specificity in new ligand and mutant protein structures.*
- Designer enzymes and small molecule recognition materials play an important role in the science of: WMD sensing (thrust 1), protection (thrust 3), and securing WMD (thrust 5)
- More broadly it is important in pharmaceutical drug target identification and development, and enzyme engineering for applications such as bioenergy

Background and Significance

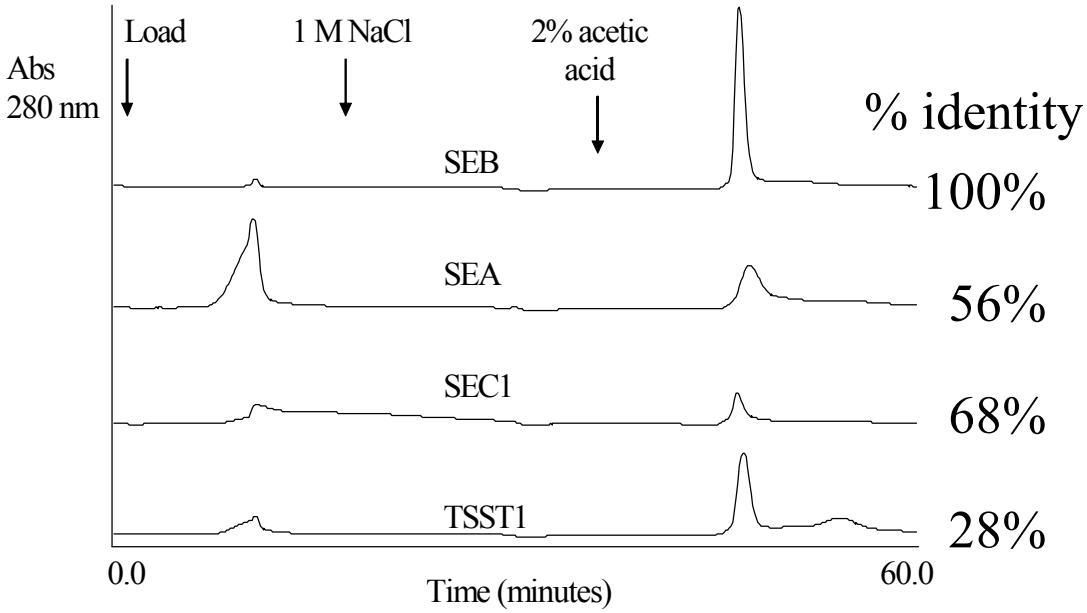
WMD Sensing: Conservation of Structural and Functional Motifs

Practical Problem: Replace antibodies in detection assays with ligands that have “guaranteed” defined species specificity.

Solution: Make ligands that bind to structural features of proteins that are evolutionarily conserved across a given species or functional class, but not shared with other species.

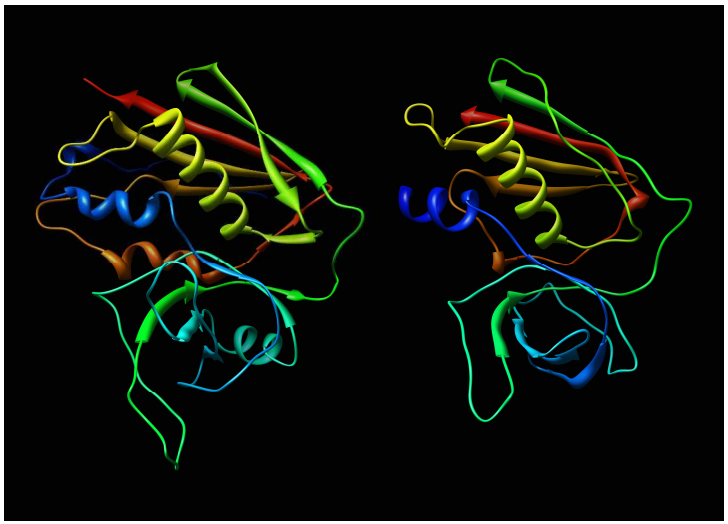
Sequence similarity does not predict binding affinity

Example 1: SEB-binding peptide



Wang, G., De, J. Schoeniger, J.S., Roe, D.C. and Carbonell, R.G. (2004) *Journal of peptide research* **64,** 51-64

SEB and TSST are structurally homologous but have low sequence homology (28%)



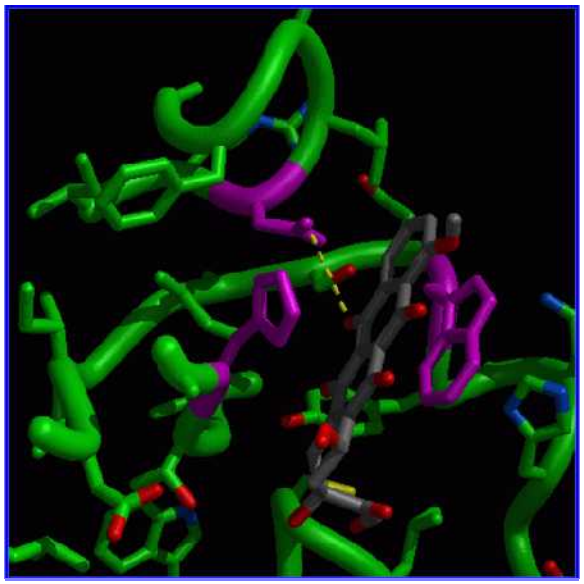
SEB TSST

Sequence similarity does not predict binding affinity

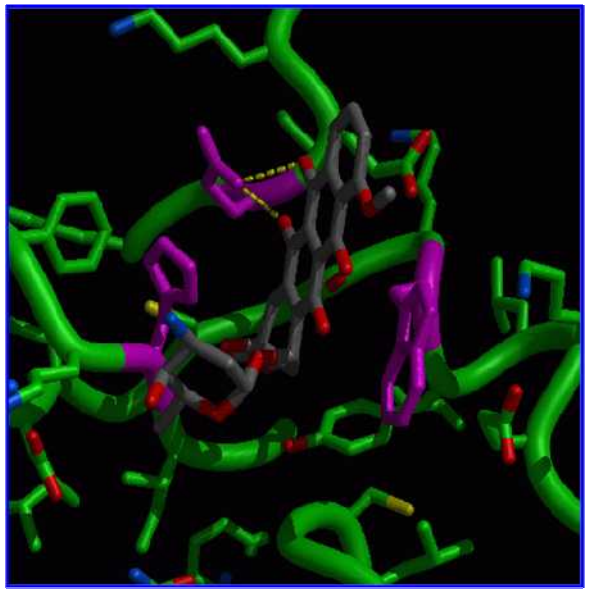
Ex 2: Docked Molecule binds to botulinum and tetanus toxins

- Virtual screen against tetanus toxin
- 15 (out of ~30 tested) confirmed experimentally
- Top compound (doxorubicin) bound to BoNT B as well (38% identity)

Dock-predicted binding of doxorubicin to TeNT



Crystal structure binding to BoNT(S. Swaminathan)



Chem. Res. Toxicol. (2002), 15(10), 1218-1228.
Chem Res Toxicol **13**(5):356-62. 2000.

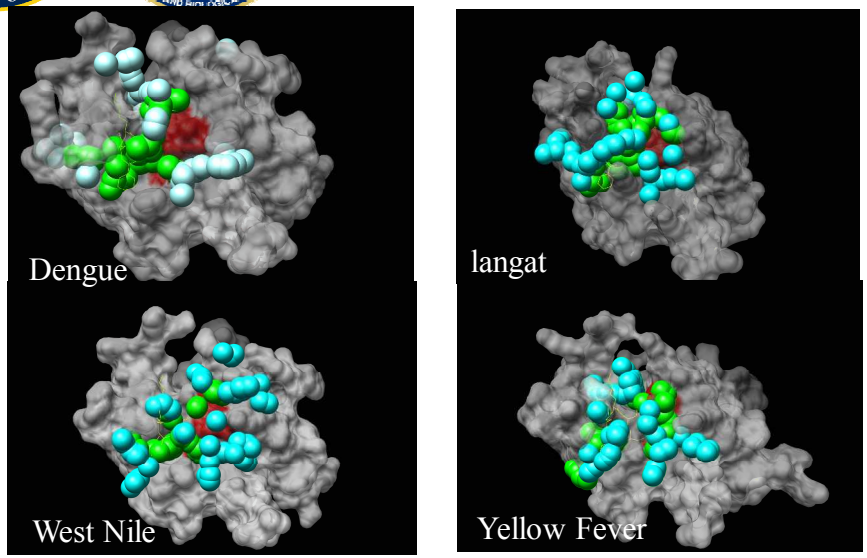
Background and Significance

Countermeasures: Conserved Binding Motifs

Practical Problem: For a given set of target organisms and target proteins, determine when it might be possible discover a drug candidate that has broad spectrum activity against the class, or which subsets might be logical co-targets.

Solution: Classify proteins based on their potential ligand interactions.

Clustering Using Ligand Binding Profiles from Docking



Clustering by MSA Yields 4 Groups

Common motifs

Catalytic Triad

Other Active Site

In silico binding scores were found for a test set of 1000 diverse molecules docked as ligands to modeled structures. Sequences were re-clustered based on cross-correlation of scores

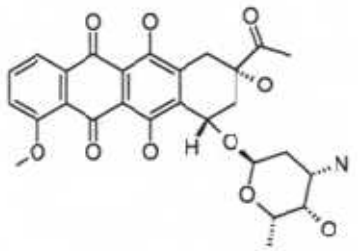
		correlated docking score with representative structure				
	sequence	dengue	WNV	langat	YF	Modoc
cluster1	WNV_rot	0.863	1	0.602	0.606	0.786
	Kunjin_6695_12.2ijo_B	0.864	0.98	0.595	0.602	0.785
	WN_6695_9.2ijo_B	0.864	0.968	0.587	0.597	0.782
	Zika_6695_8.2ijo_B	0.874	0.878	0.718	0.706	0.756
	Alfuy_6695_15.2ijo_B	0.838	0.876	0.52	0.526	0.759
	MVE_6695_17.2ijo_B	0.829	0.871	0.503	0.502	0.757
	2fom_rot	1	0.863	0.627	0.628	
	Usutu_6695_19.2ijo_B	0.793	0.859	0.42	0.42	0.752
	Kedougou_6695_6.2ijo_B	0.885	0.852	0.697	0.691	0.751
	Ilheus_6695_7.2ijo_B	0.8	0.836	0.674	0.682	0.708
	SLE_6695_14.2ijo_B	0.785	0.832	0.518	0.532	0.732
	Dengue4_6695_4.2ijo_B	0.847	0.814	0.626	0.578	0.702
	Rocio_6695_10.2ijo_B	0.781	0.814	0.73	0.7	0.691
	Dengue1_6695_2.2ijo_B	0.877	0.809	0.582	0.575	0.706
	Dengue3_6695_3.2ijo_B	0.896	0.809	0.744	0.741	0.72
	Yokose_6695_21.2ijo_B	0.8	0.76	0.789	0.774	0.701
cluster2	YF_rot	0.628	0.606	NA	1	NA
	langat_rot	0.627	0.602	1	0.849	0.612
	Omsk_6695_26_2snv.2fom	0.628	0.596	0.975	0.849	0.609
	TBE_6695_27_2snv.2fom_I	0.624	0.59	0.975	0.853	0.603
	LoupingIll_6695_29_1df9_A	0.624	0.598	0.956	0.849	0.605
	Entebbebat_6695_20.2ijo_I	0.719	0.691	0.869	0.863	0.631
	Sepik_6695_5.2ijo_B	0.7	0.677	0.863	0.867	0.603
	Karshi_6695_28_1df9_A.2f	0.693	0.687	0.846	0.769	0.712
	JEE_6695_18.2ijo_B	0.764	0.756	0.818	0.832	0.683
	RioBravo_6695_22.2ggv_B	0.68	0.673	0.806	0.883	0.635
cluster3	Modoc_6695_24_2snv.2fon	0.773	0.786	0.612	0.581	1
	MontanaMyotisLeukoE_669	0.721	0.721	0.728	0.697	0.846
outliers	Powassan_6695_30_1qy6_	0.251	0.298	NA	0.22	0.432
	Bagaza_6695_13.2ijo_B	0.52	0.591	0.117	0.122	0.494

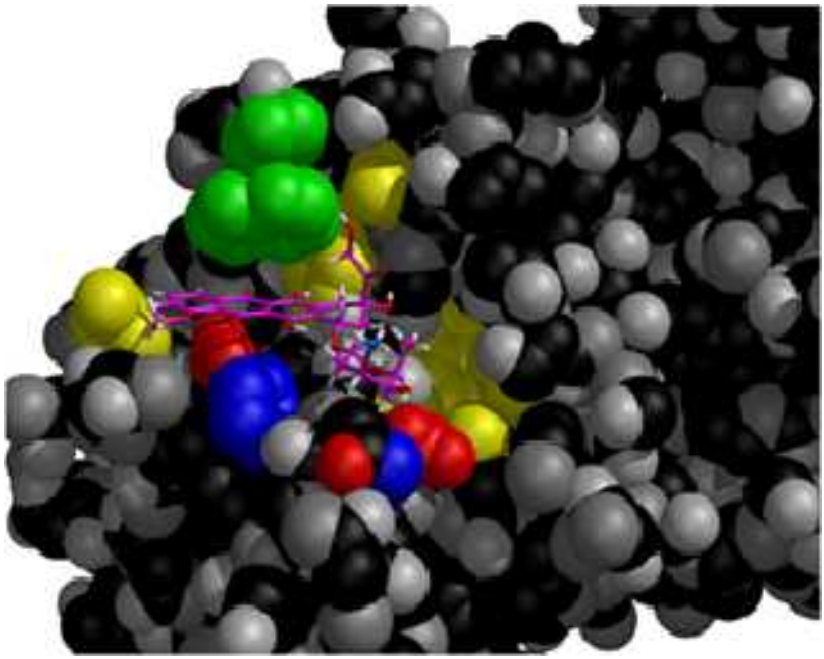
Clustering by Ligand Binding Yields 3 Groups, with Dengue & WNV now together in one group

Background and Significance

What is a binding motif?

- Interface between protein (macromolecule) & small molecule
  - enzyme-substrate, antibody-antigen, drug-target, etc.
- Mathematically ill-defined (as a geometric entity) for purposes of clustering
  - Surface painted with scalars & vectors
  - Actually dynamic (Non-rigid geometry)
- Must be analyzed across a vast space of ligands and receptors





Technical Approach

- **GOAL: Find specificity-determining features (SDFs) across protein target and ligand spaces**
- Identify Promising Target Families
  - Lots of protein variants known, lots of ligand data available
  - Applications potential: Primarily infection & immunity (drug targets)

Test System	Enzyme Source	Experimental Ligand Binding Data Available in literature
Protein Kinases	Human	>40,000
DHFR	Bacterial / fungal / protist	> 4000
HIV / HCV Proteases	Viral	>14,000 / >300

Find specificity determining features (SDFs)

Binding Data (Ki, Kd)

Ligands

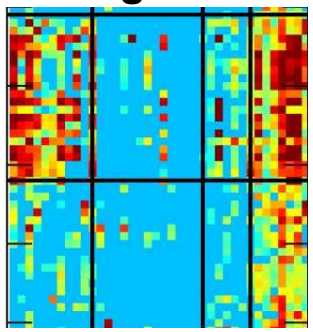
Proteins

Table of Binding Data

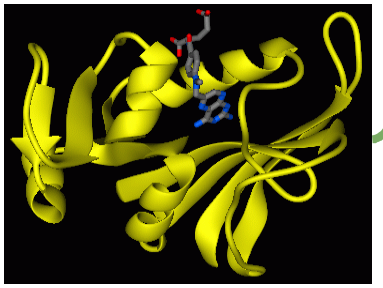
Classify

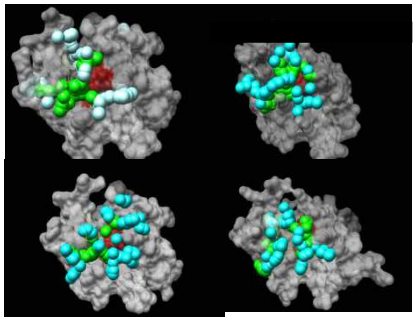
Ligands

Proteins

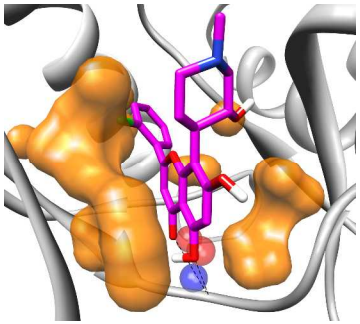


Simulation (Docking, MD)

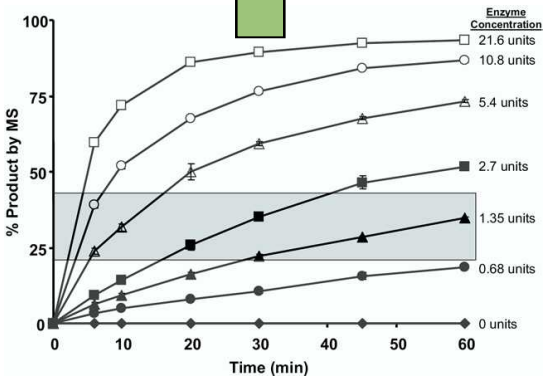




Extract SDFs



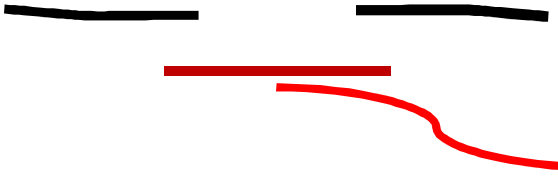
New predictions



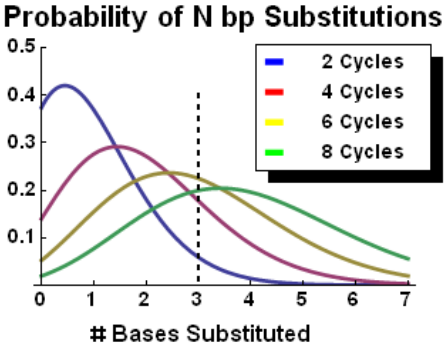
Experimental Validation

Generating experimental binding data for ALL Triple Mutants of Binding Site

PCR out gene segment around the binding site (~ 70 bp)



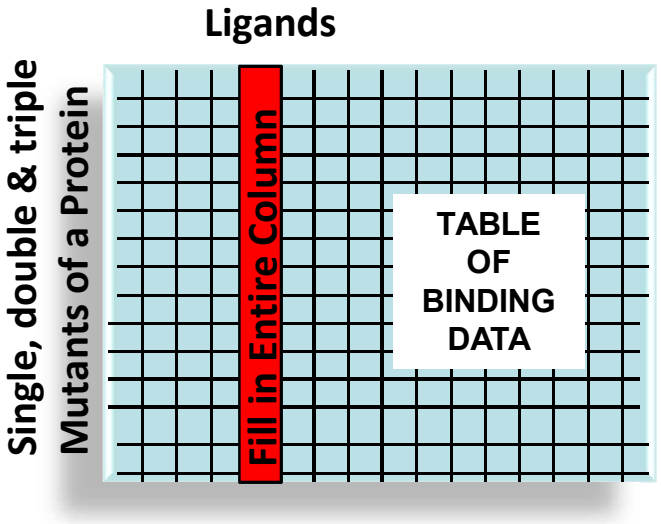
Amplify with Error-prone PCR to get all single, double and triple mutants



Reintegrate into gene & Express in phage protein display system (<= 1 protein copy per phage) to produce a library of triple (bp not aa) mutants.

Affinity chromatography versus immobilized ligands sorts out weak, medium and strong binders

Illumina Sequencing of each fraction determines which mutants are in it. 1-2 Million reads (~$200) provides >4x coverage of ALL triple mutants. Since coverage is complete, can be repeated for additional ligands.



Technical Approach

Benefit and uniqueness of SDF Approach

- Classifying across **both** protein and ligand space
  - Can include all interesting ligands and targets (ex: off-target receptors)
  - Can incorporate other data (e.g. toxicity)
- Can include whatever binding data available. Clustering sorts out weighting for you
- Framework for integrating computational & experimental
  - Not depending on high fidelity simulations
  - Allows statistical analysis

Technical scope and limitations

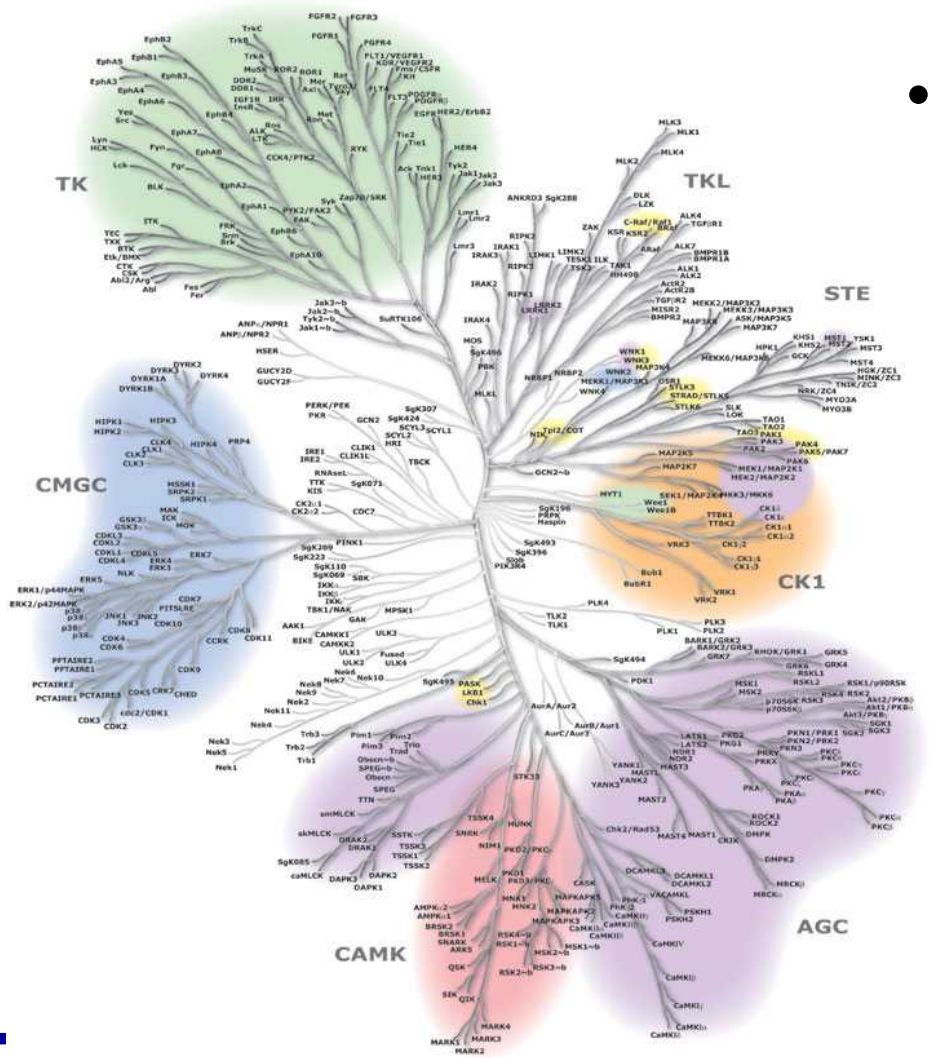
- Intelligently sampling protein/ligand space
  - $>10^{60}$ possible small molecules
  - $>20^{300}$ possible proteins
  - Ligands: choose diverse, biologically relevant, commercially available
  - Proteins: Families within a species, homologs, mutations

Strategy/Risk Mitigation

Difficulties:

- Hard to get complete data set for large protein/ligand space. Lots of "missing" data
  - *Missing data sensitivity analysis*
  - *Selectively fill in experimentally*
  - ***Phage display to generate all triple mutants***
- Hard to get accurate simulations of binding data (real structures are dynamic)
  - *High-fidelity: Perform selective high-fidelity simulations and use information for related systems*
  - *Low-fidelity: Perform simulations on data sets large enough for statistical analysis*

Research Progress

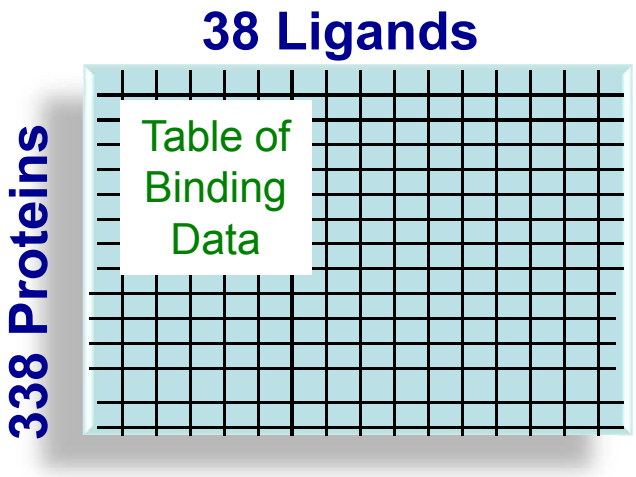


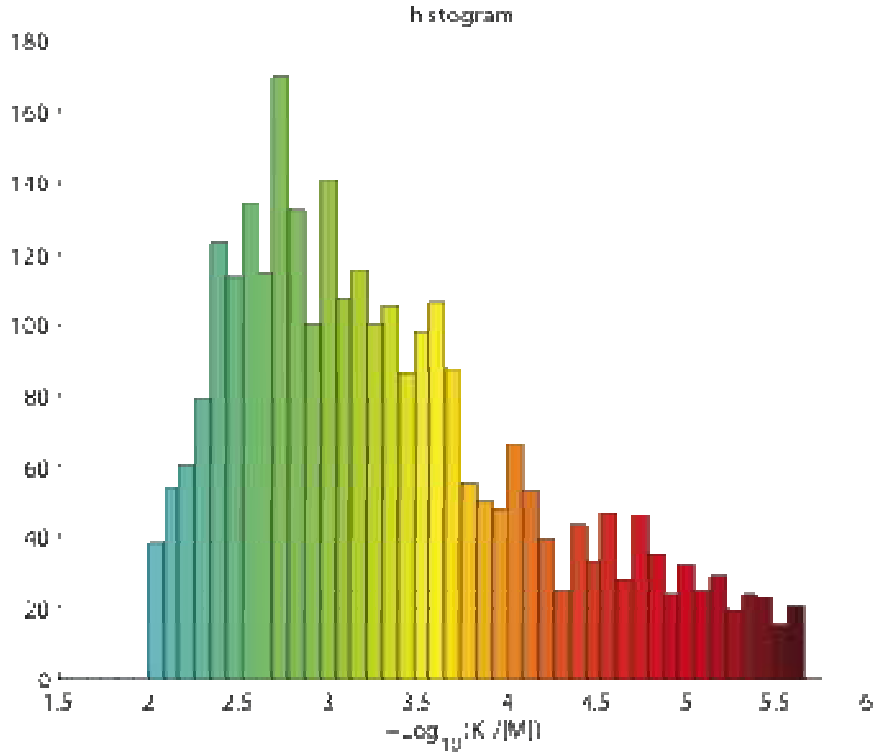
- **The human kinome**
  - 40 atypical PKs
  - 478 classical PKs.
    - 388 serine/threonine kinases,
    - 90 tyrosine kinases
    - 50 sequences which lack a functional catalytic sites.

Manning et al., Science, 6 December 2002

Results: Starting TBD for the human kinome

Values for Kinase/Ligand TBD taken from a comprehensive experimental study in the literature.



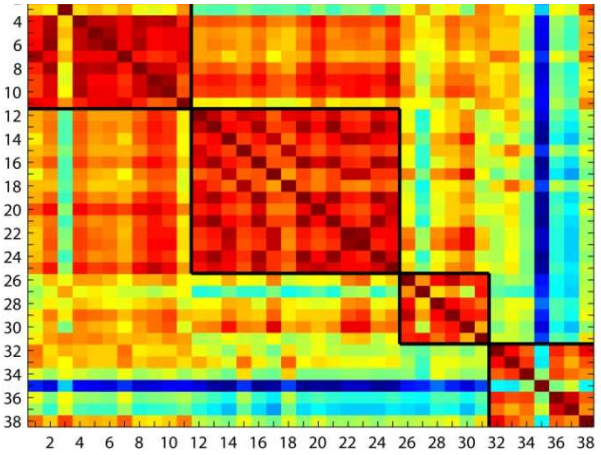


Karaman MW, et al, Nat. Biotechnol, 2008. ***26*** *127-132.*

Human Kinome Results: Cluster by ligand binding data

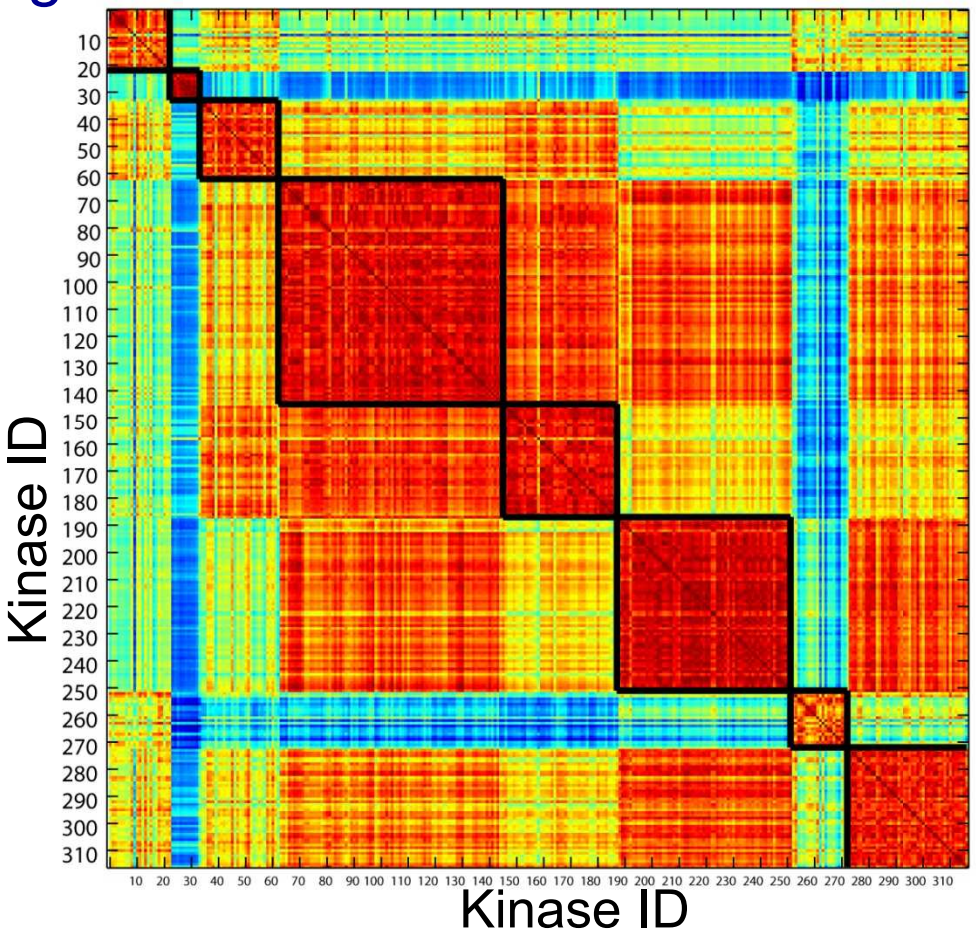
Ordered Heatmap showing kcenters clusterings

Ligand Clustering



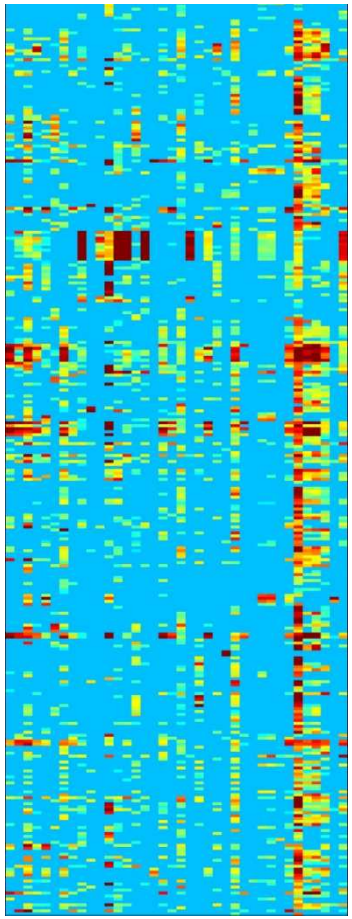
- All “type-2” inhibitors in ligand cluster 1
- All broad binders in ligand cluster 4

Protein Clustering

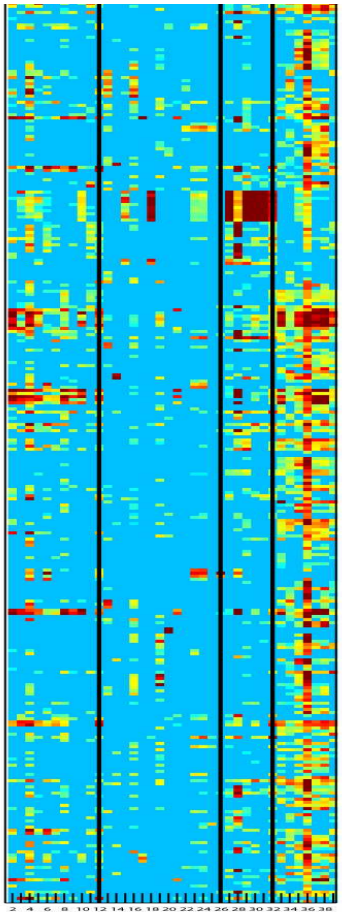


Human Kinome Results: Binding Data Ordered by Clusters

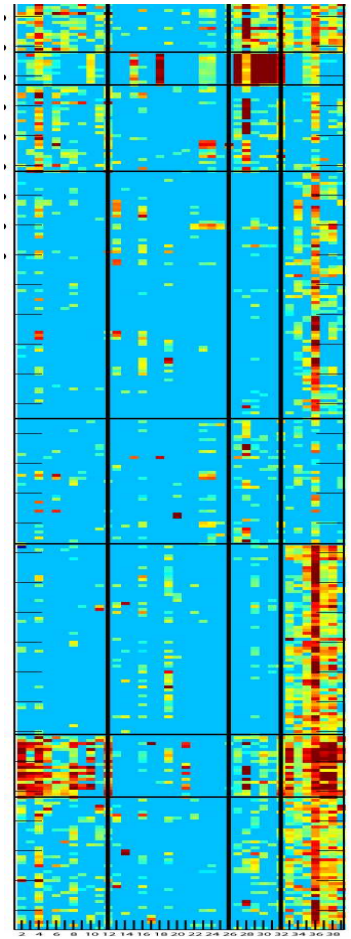
Unordered Binding Matrix



Binding Matrix Ordered by Ligand Clusters



Binding Matrix Ordered by Ligand and Protein Clusters

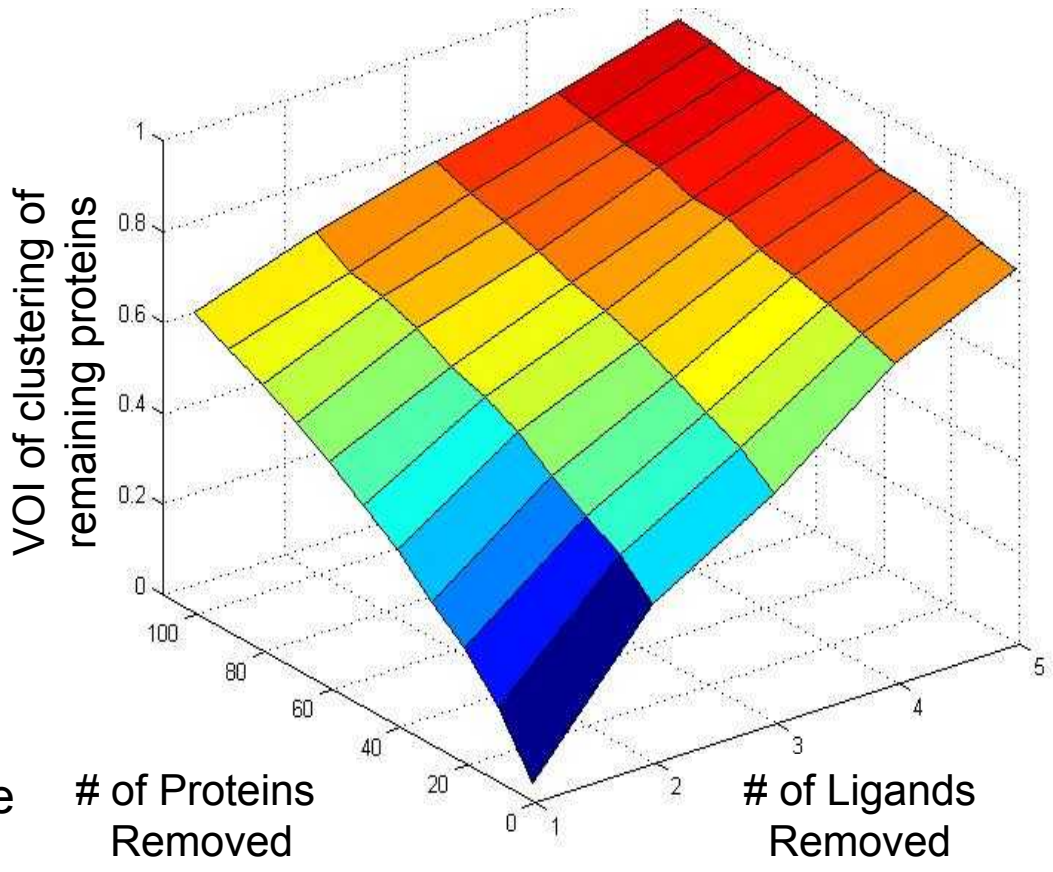


Robustness of Classifications

Leave 1-out analysis shows clustering robust for both ligands and proteins

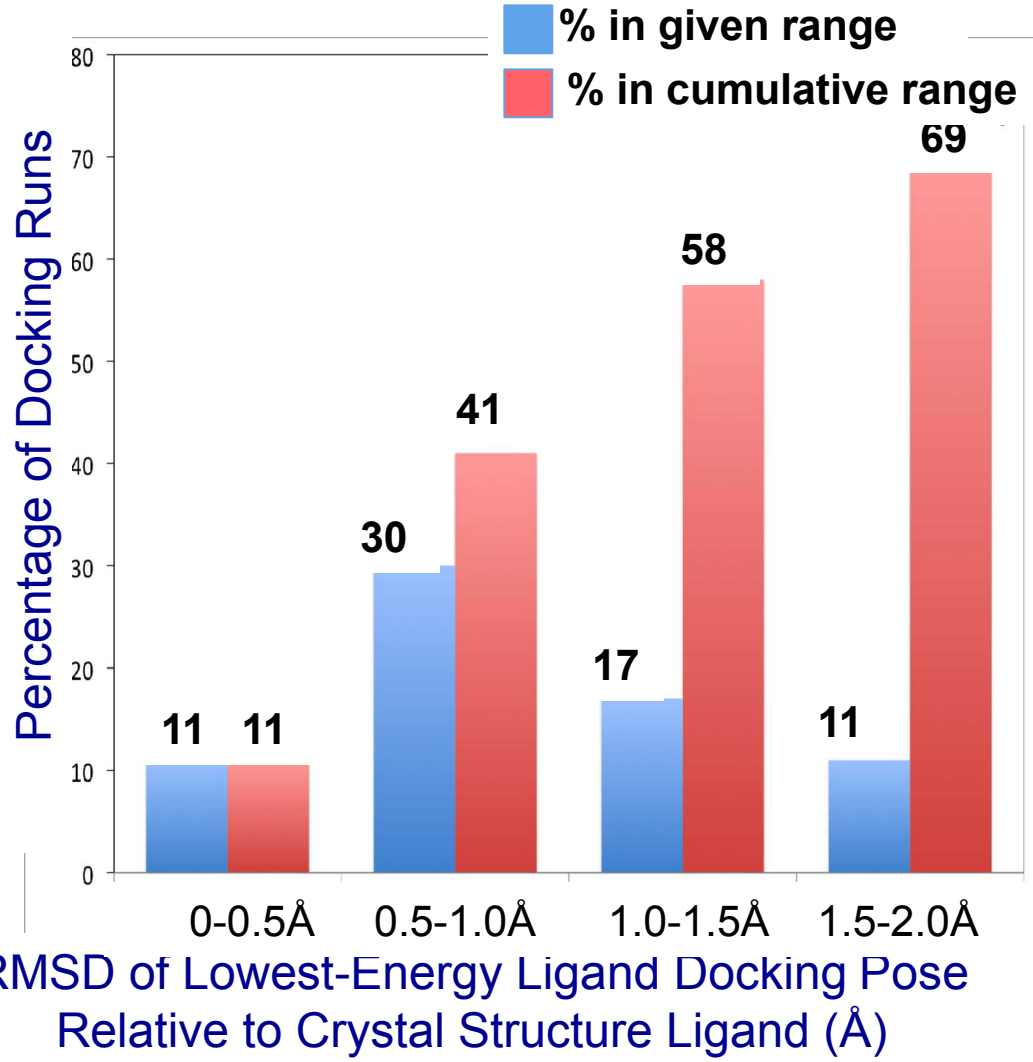
- Variation of information (VOI) Mathematical method to measure distance between 2 clusterings.
- *Clustering by sequence or structure do not capture the patterns in experimental data.*
  - VOI of random cluster is 3.7
  - VOI for clustering by sequences is 2.57
  - VOI for clustering by structure motifs is 2.73

Cluster Degradation with respect to protein and ligand removal



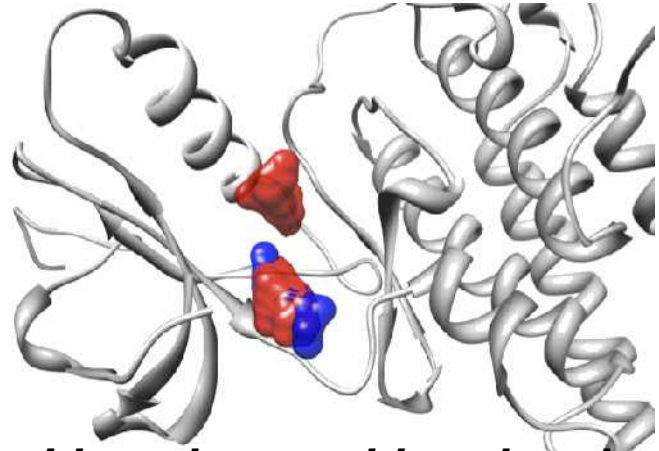
Docking to kinases and extracting specificity determining features (SDFs)

- Docked 38 ligands to 113 kinase structures using autodock 4 with flexible ligands
- Validated docking poses with crystallographic ones for those with co-crystals (figure)
- Features (h-bonds, polar, hydrophobic) extracted from docked poses using experimentally determined clusterings.
- Statistical approach to feature extraction– insenstive to “noise” from mis-docked features

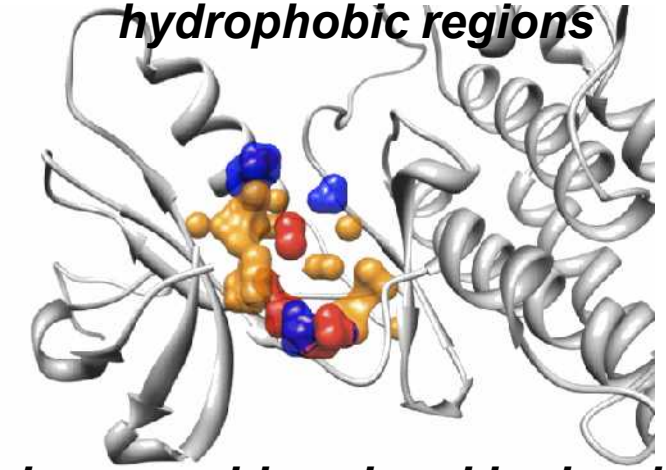


SDFs: Broad Binding Features (common among all clusters)

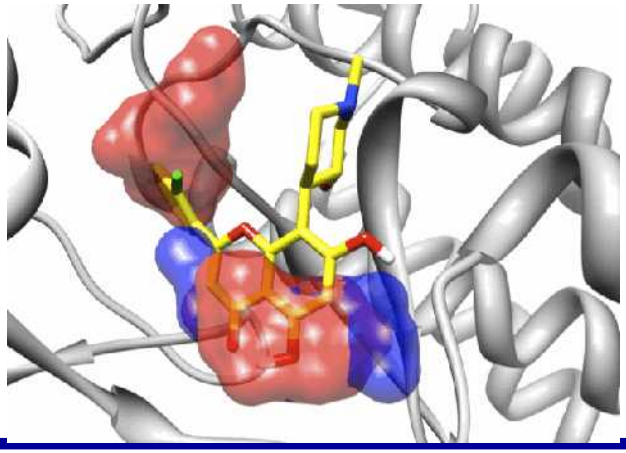
Ligand-space hbond regions



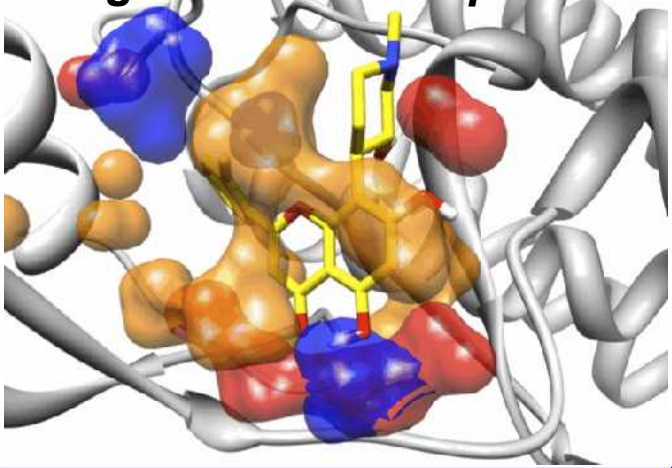
Protein-space hbond and hydrophobic regions



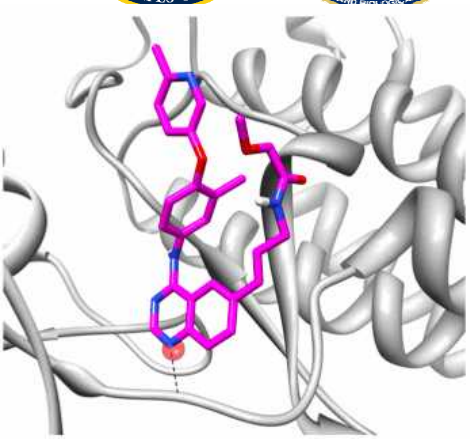
Ligand-space hbond regions with flavopiridol



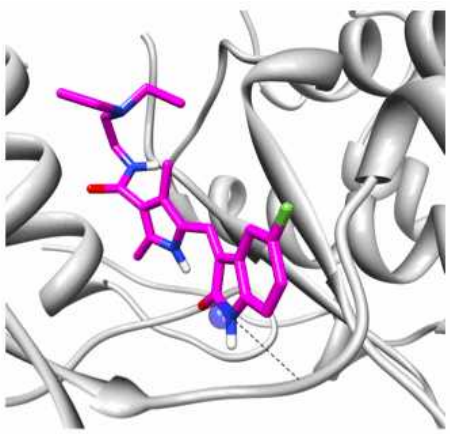
Protein-space hbond and hydrophobic regions with flavopiridol



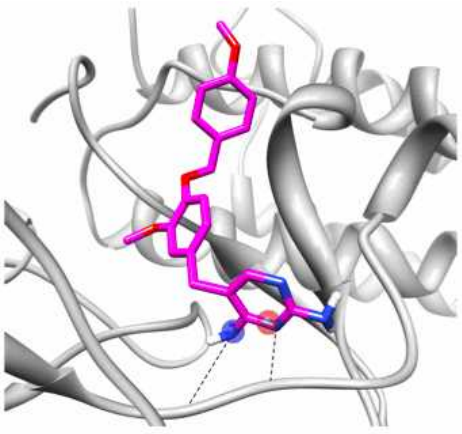
Docked Conformations Agree With Extracted SDFS



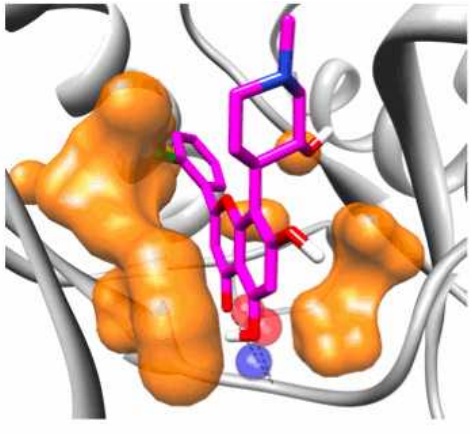
ZAP70, CP-724714



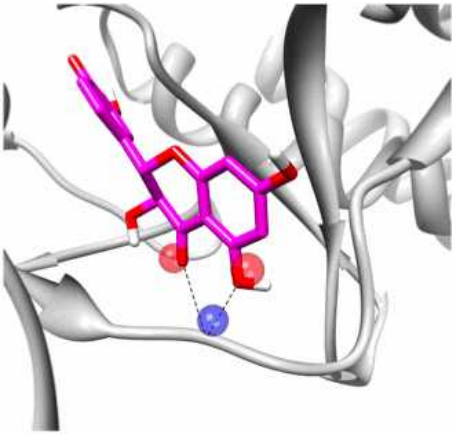
JNK1, sunitinib



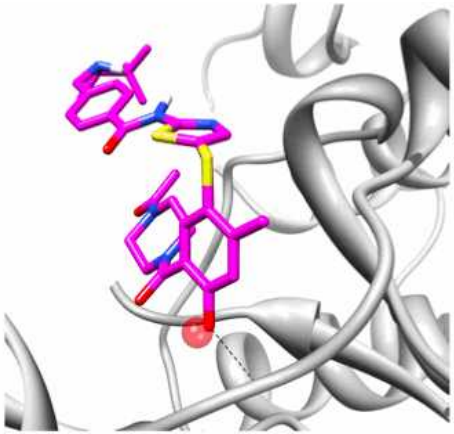
LYN, GW-2580



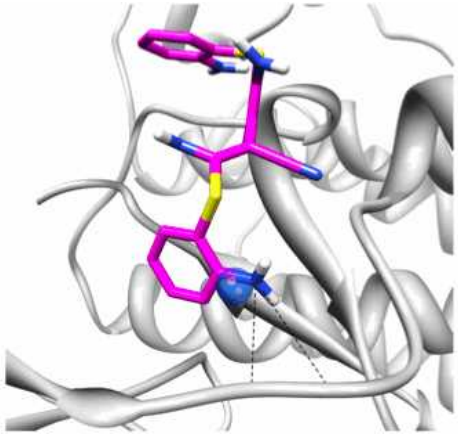
CLK1, flavopiridol



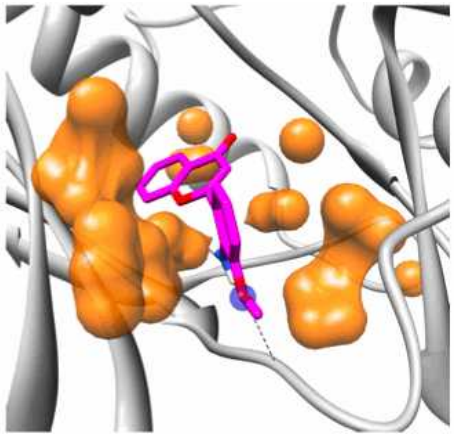
ITK, quercetin



MEK6, BMS-509744

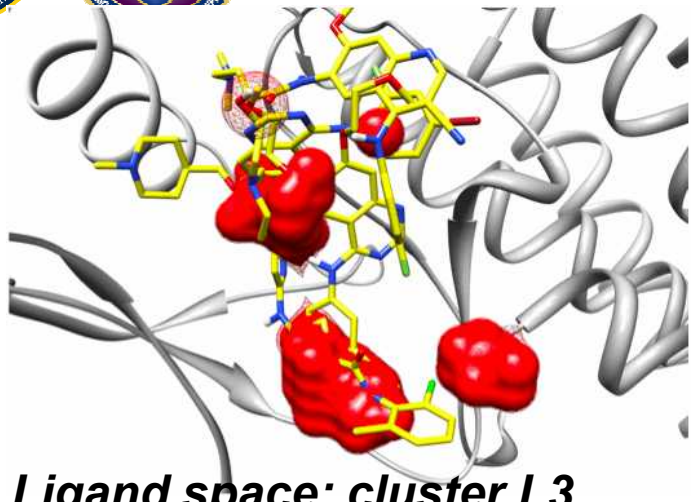


ITK, U0126

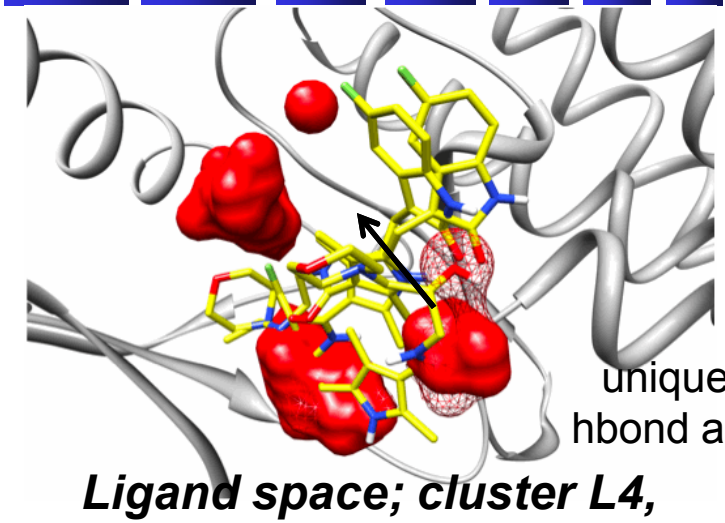


KIT, PD98059

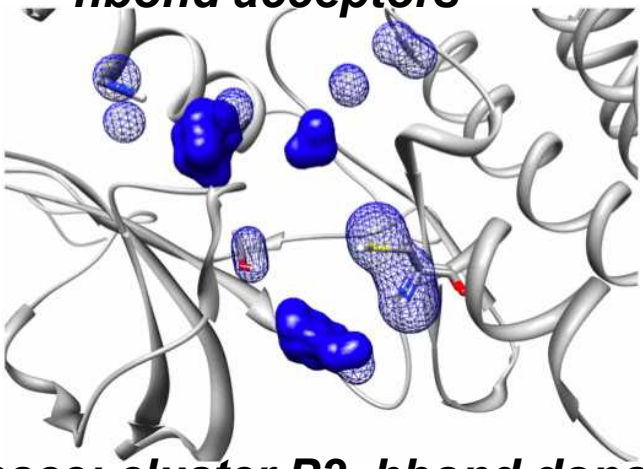
SDFs Unique to a Cluster



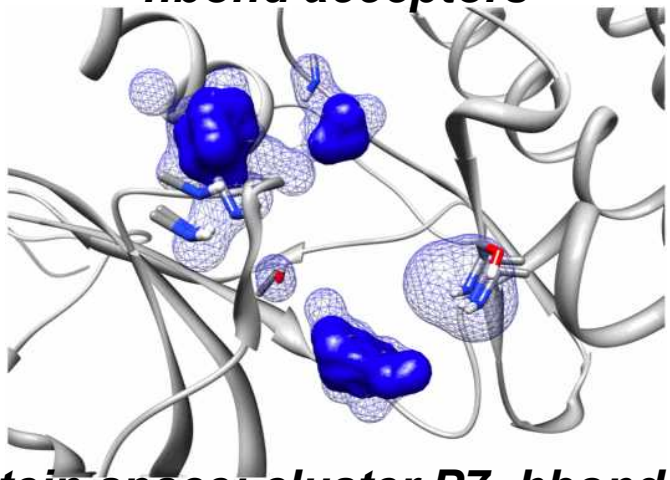
Ligand space; cluster L3, hbond acceptors



Ligand space; cluster L4, hbond acceptors



Protein space; cluster P2, hbond donors



Protein space; cluster P7, hbond donors

Summary of Kinase Study

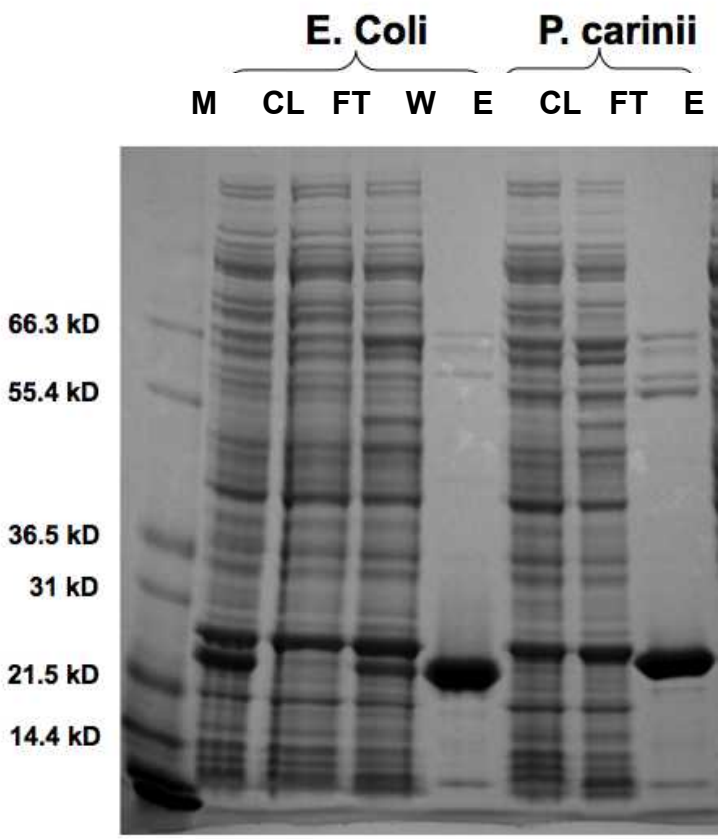
- Using ligand binding data is a robust way to cluster proteins and ligands and useful patterns of binding emerge from these clusterings.
- We can turn combine these clusters with docked poses to extract SDFs
- These SDFs match specificity features in ligands outside our initial data set.
- Next step: experimental validation

Experimental Progress: Proteins

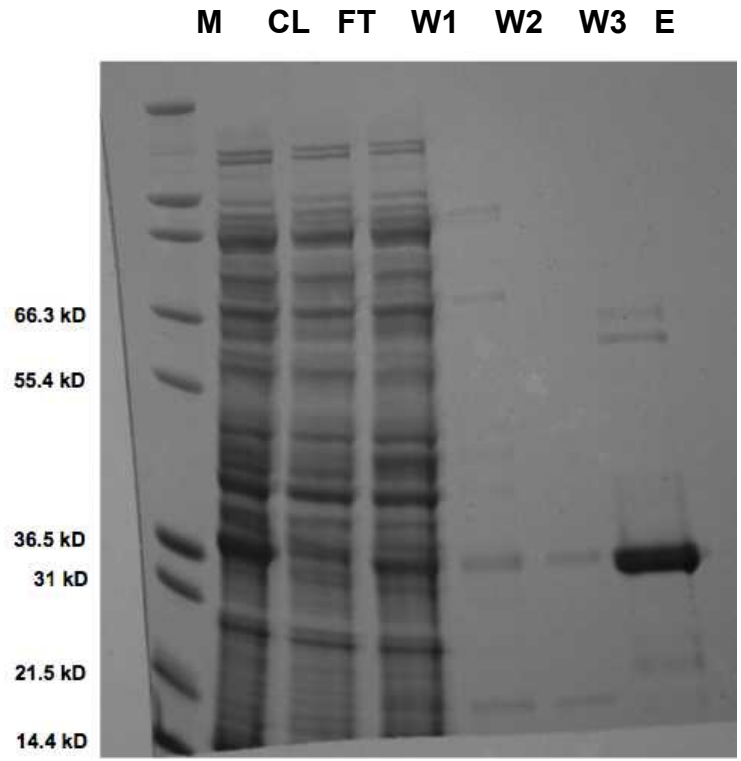
- Kinases: Several Commercially Available Purchased
- DHFR
  - High yield of active DHFRs expressed in *E. coli*
    - *E. coli* DHFR : 42 mg from 250 ml culture
    - *P.carinii* DHFR: 33 mg from 250 ml culture
  - Active DHFRs displayed on T7 phages
- HCV protease
  - Constructed HCV NS3 protease and NS4A cofactor peptide as a single-chain
  - High yield of the active protein: expressed in E. coli with Sumo tag
- HIV Protease
  - Expressed in E. coli with Sumo tag gave high yield but not active
  - Need to refold the protein from inclusion body

High yield of pure target proteins

DHFR



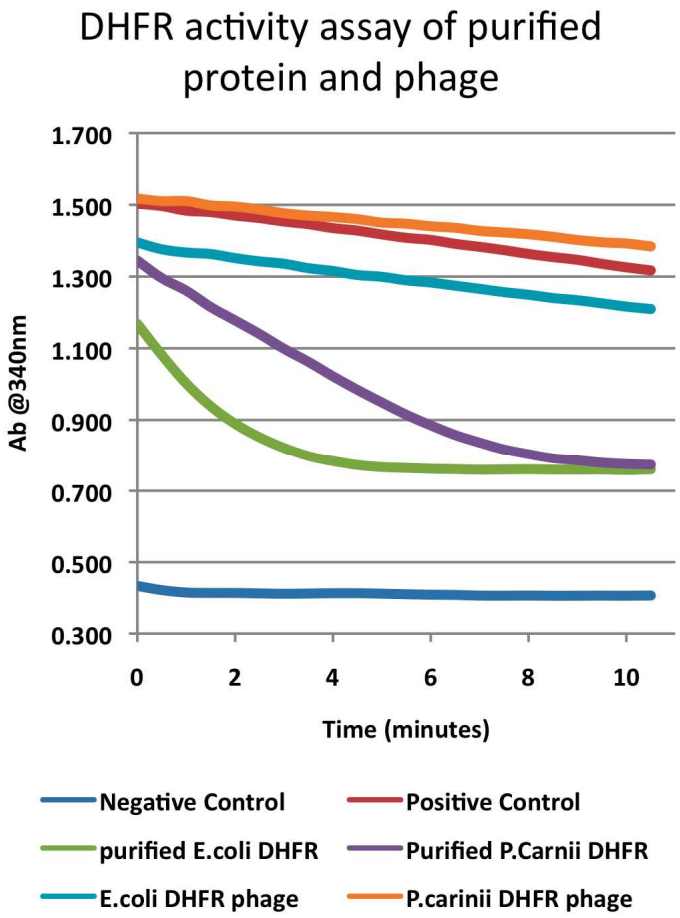
Sumo-HCV



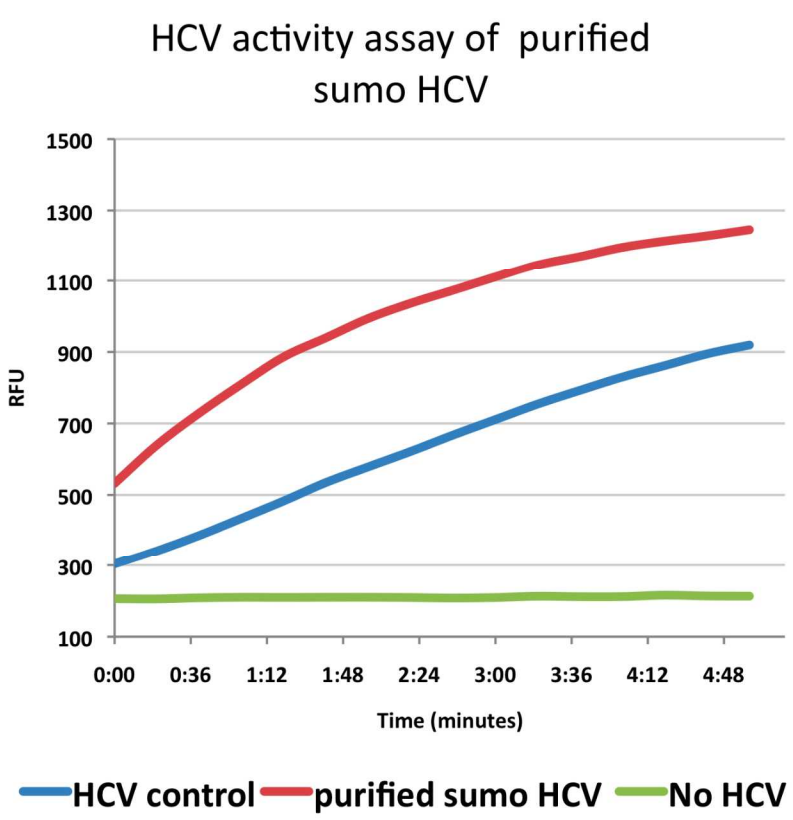
M: MARK12, CL: cell lysate, FT: flow through, W: wash, and E: elution fractions

The expressed target proteins are active

DHFR



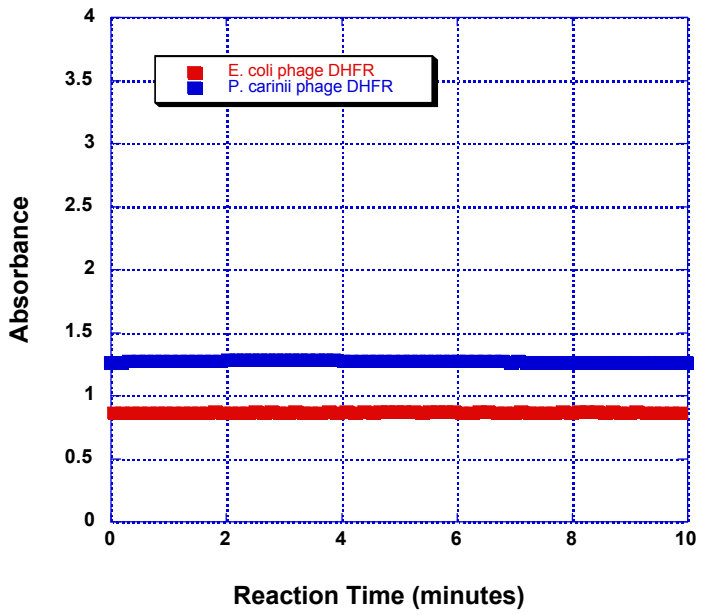
HCV



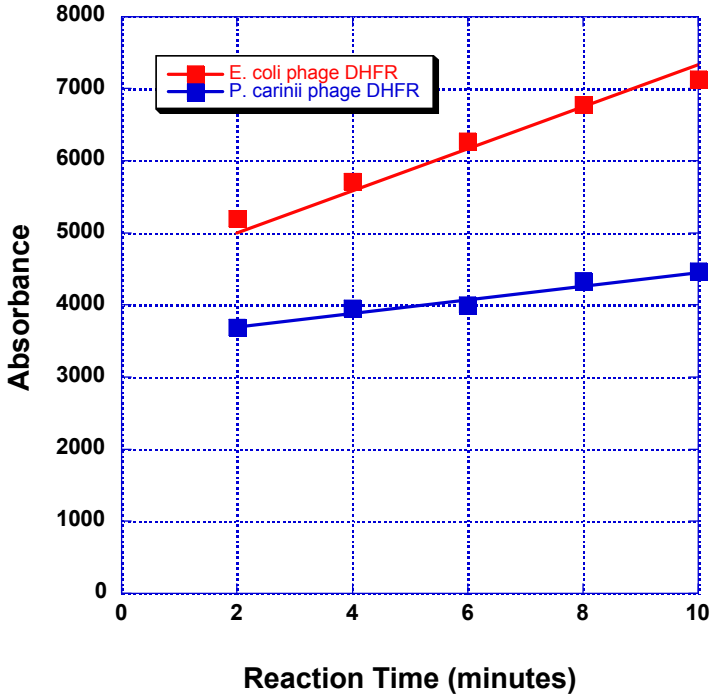
We implemented an Ultrasensitive DHFR Activity Assay

Typical activity assay monitors DHFR ability to catalyze the reversible NADPH-dependent reduction of DHF to THF

Standard assay: DHF depletion by absorbance at 340 nm.
No perceivable change



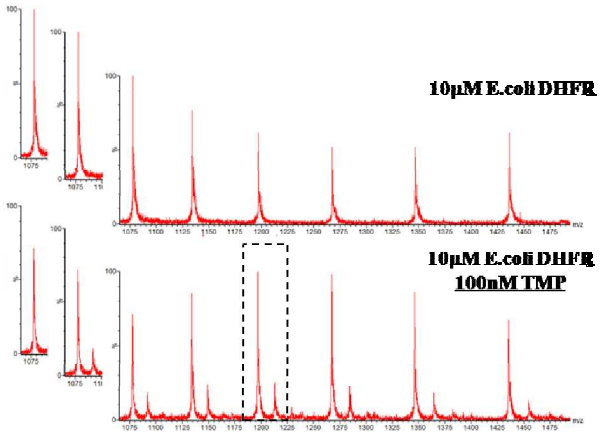
Improved sensitivity: monitor THF formation .
Activity detected



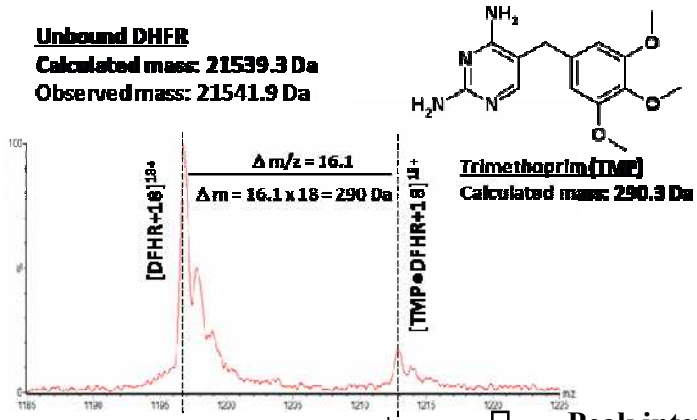
Experimental Progress: Ligands

- Kinases:
  - purchased several commercially available ligands
  - Mass Spec-based activity assay identified
- DHFR
  - purchased several commercially available ligands
  - Mass Spec-base binding Screening method implemented
- Viral Proteases
  - Sensitive Fluorogenic substrate-based activity assay implemented in micro-titer plates

Mass Spec of Non-Covalent Complexes for Measuring Kd

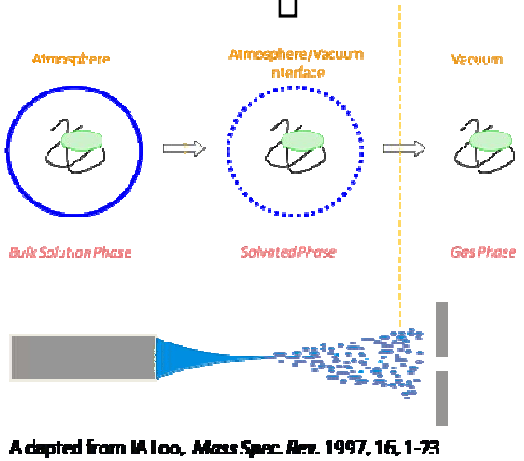


Direct observation of free and bound substrate



Trimethoprim

LIGANDS



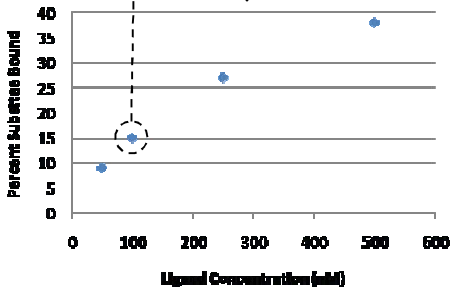
Adapted from M Ioo, *Mass Spec. Rev.* 1997, 16, 1-73

E. Coli DHFR

PROTEINS

TABLE OF BINDING DATA

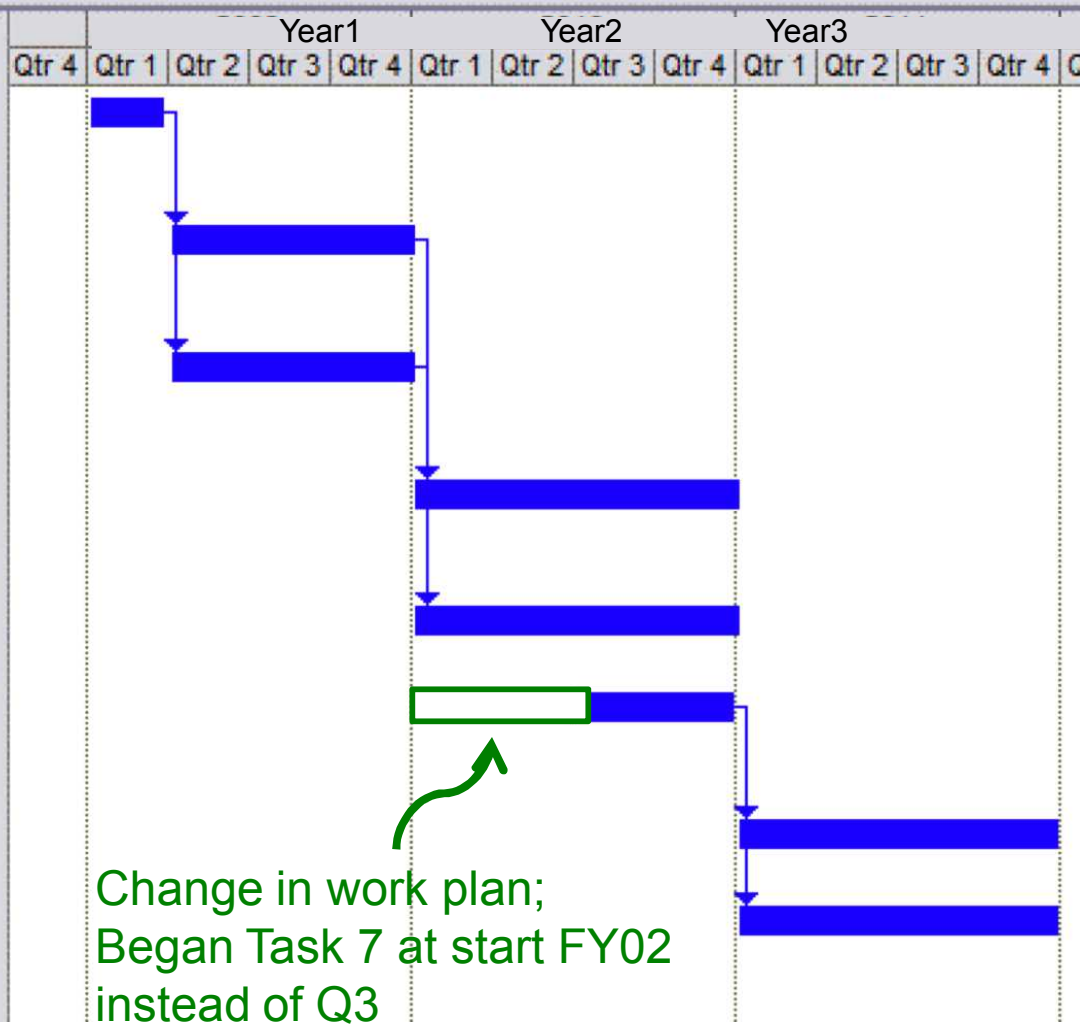
Peak intensities correlate to bound:free ratio



Kd = 10^{-9} Molar

Program Lifecycle

Task Name
Choose appropriate model experimental systems, each with a known range of specificity within ligand set and protein partners.
Evaluate predictive capability of sequence and structure-based protein classification with experimental binding specificity from initial model system data (control)
Validate predictive capability of CCM protein classification with experimental binding specificity using initial model system data.
Evaluate predictive capability of sequence and structure-based protein classification with experimental binding specificity using full data set. (control)
Validate predictive capability of CCM protein classification with experimental binding specificity using full data set.
Generate and test multiple hypotheses from CCM feature identification for physical and chemical mechanisms in ligand and protein binding specificity
Validate CCM predicted ligands with tailored molecular recognition specificity
Validate CCM predicted proteins with tailored molecular recognition specificity using mutant assays.



Conclusions

- **This study supports our hypothesis that studying protein/ligand binding can provide more insight into SDFs than structure or sequence alone**
  - Classifications based on structure/sequence lose information
  - Dual treatment of ligands &proteins enables features of both that contribute to specificity to be extracted.
- **This study provides multiple new hypotheses that we can test experimentally:**
  - Hypotheses for features that determine broad and narrow binding ***within*** a protein family
    - We can add new features such as protein dynamics/water interactions and test them
    - We can test for features that may cause drug resistance
  - We can also test the ability of SDF models for features **between** different protein families to categorize unknown enzymes

Project Deliverables

- Publications: 2 journal articles in preparation
- Presentations:

  "Conserved Motifs to Examine the Effects of Sequence Variation in Pharmaceutical Chemical and Biological Defense Science and Technology Conference, Nov. 16-20, Dallas, TX. 9292, Livermore, CA 94551

  "Classifying proteins by common, conserved motifs"; ACS Spring Meeting March 21-25

- People supported (partials included):
  - 3 postdocs
  - 2 interns (undergraduate)
  - 2 technicians
  - 3 technical staff

Future Directions

- Immediate next steps:
  - SDF analysis of new test systems: DHFR, HCV/HIV
  - Experimental validation of kinases, DHFR, HCV/HIV
  - Improving our SDFs: incorporating protein dynamics, waters, ligand fragments rather than drugs;
- Further directions in basic research:
  - Correlating (SDFs) with functional pathways
  - Evolutionary Predictions
  - Enzyme Function predictions- use SDFs to categorize families and identify functions for new/unknown enzymes.
- Potential Applications
  - Countermeasure (drug) design - target selection/resistance analysis
  - Designer enzymes for protection
  - Molecular recognition materials for detection/protection

Questions

"We are all agreed that your theory is crazy. My own feeling is that it is not crazy enough."

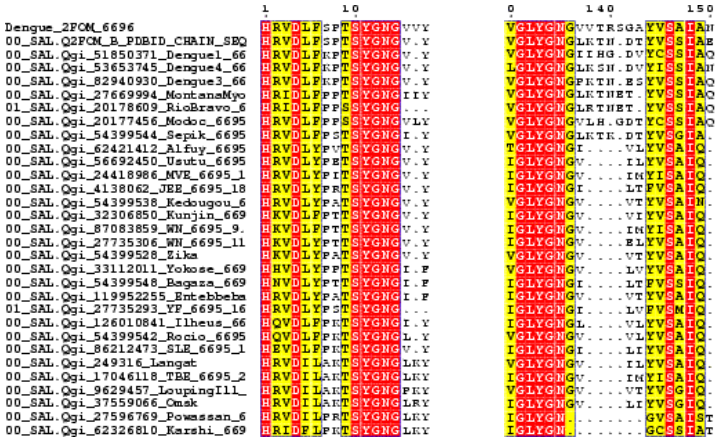
Niels Bohr

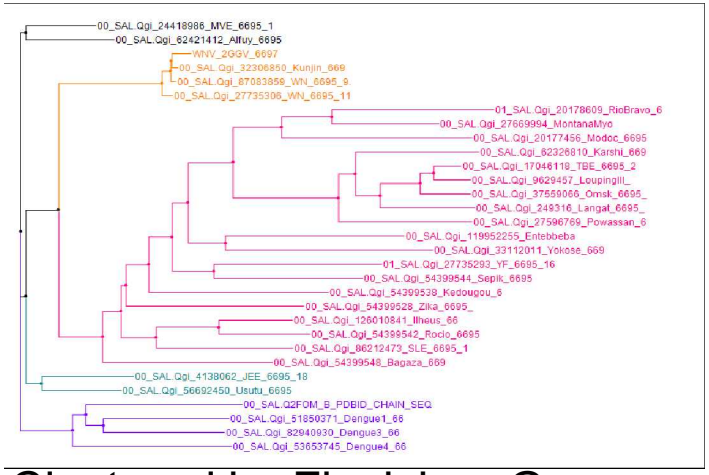
Backup Slides

Genomic or Structural Classification Reveals Four Groups of Flaviviruses

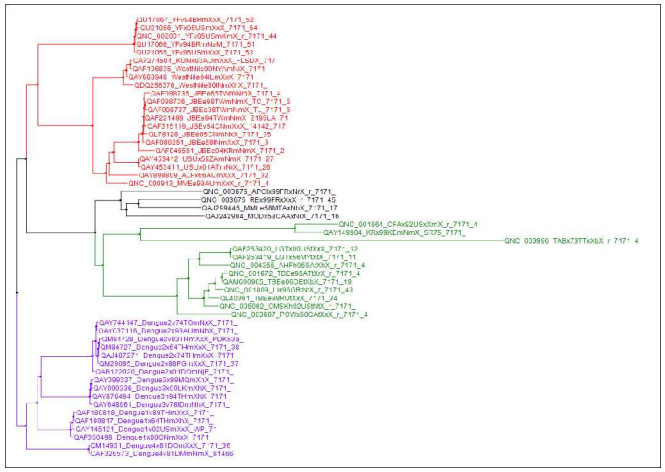
- Structural models for all sequences are aligned using substrate contacts
- Distance cutoffs to key residues define discontiguous sequence motifs
- These motifs are subjected to multiple sequence alignment

MSA of motifs close to Active Site





Clustered by Flavivirus Genome



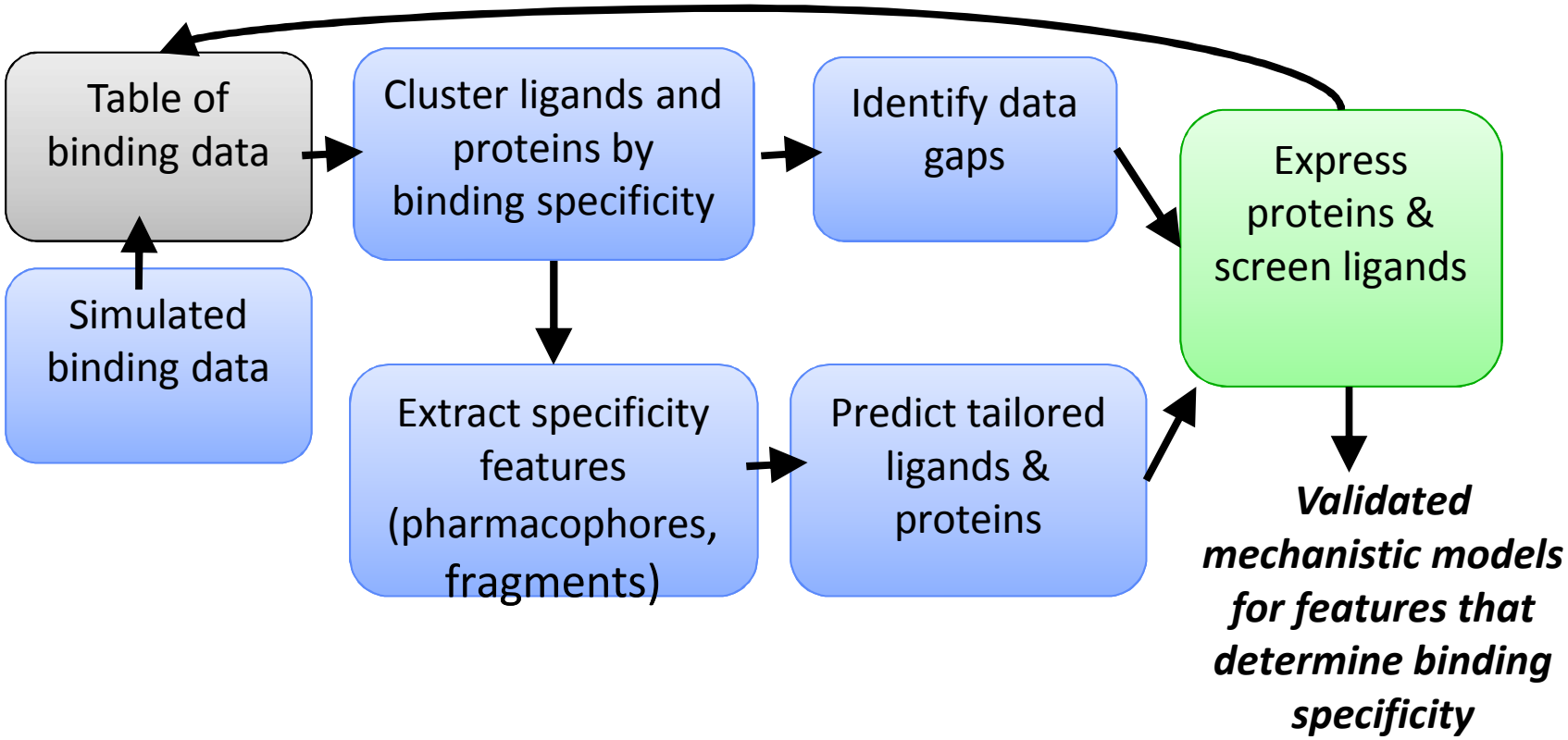
Clustered by MSA of motifs

Bad News: Dengue and WNV are never in the same group!

Technical Approach Steps

- **GOAL: Find specificity-determining features**
- Identify Promising Target Families
  - Lots of protein variants known, lots of ligand data available
  - Applications potential: Primarily infection & immunity (drug targets)
- Assemble Table of Binding Data (TBD)
  - Gather as much as possible from literature
  - Express proteins and buy ligands, test binding & fill in missing data
- Also try to simulate/predict TBD
- Use statistical methods and docking to extract features that correlate with specificity.
- Predict new binding interactions using these features
- Validate predictions on test systems

Ligand/Protein Specificity Design using CCM



Phage Display Particulars

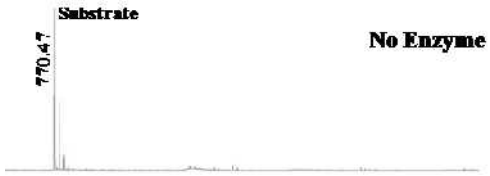
- T7 select system (Novagen)
- Protein (not peptide) display system based on bacteriophage T7
- Can control expression to display one molecule of protein per phage
  - Expression level is stochastic, so get 0.1 to 1 molecule of protein per phage on average using low level promoter

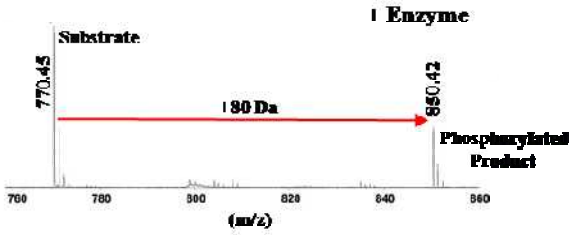
Mass Spec for Kinase Assays

Cell division protein kinase 2 (CDK2)
PDB: 3LFN

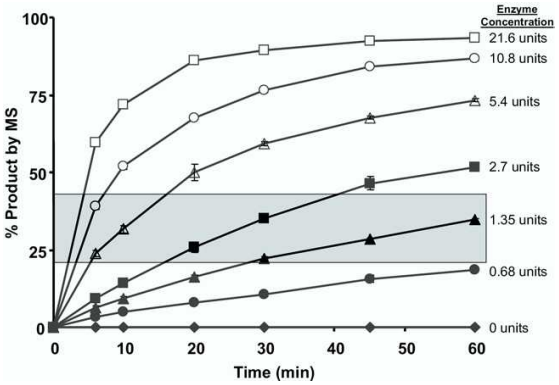
Protein tyrosine kinase 2 (PYK2)
PDB: 3ET7

Tyrosine-protein kinase (ZAP-70)
PDB: 1U59

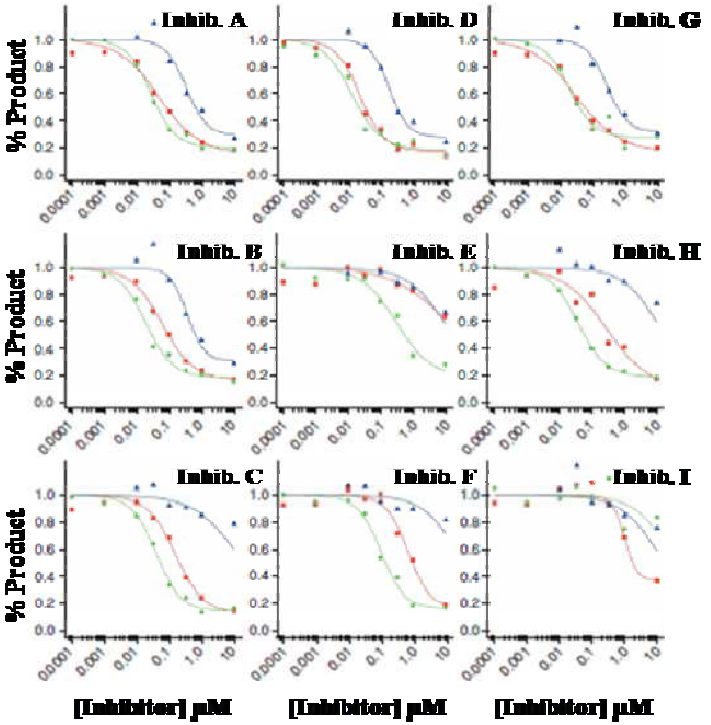




Kinase activity
in presence of ligand



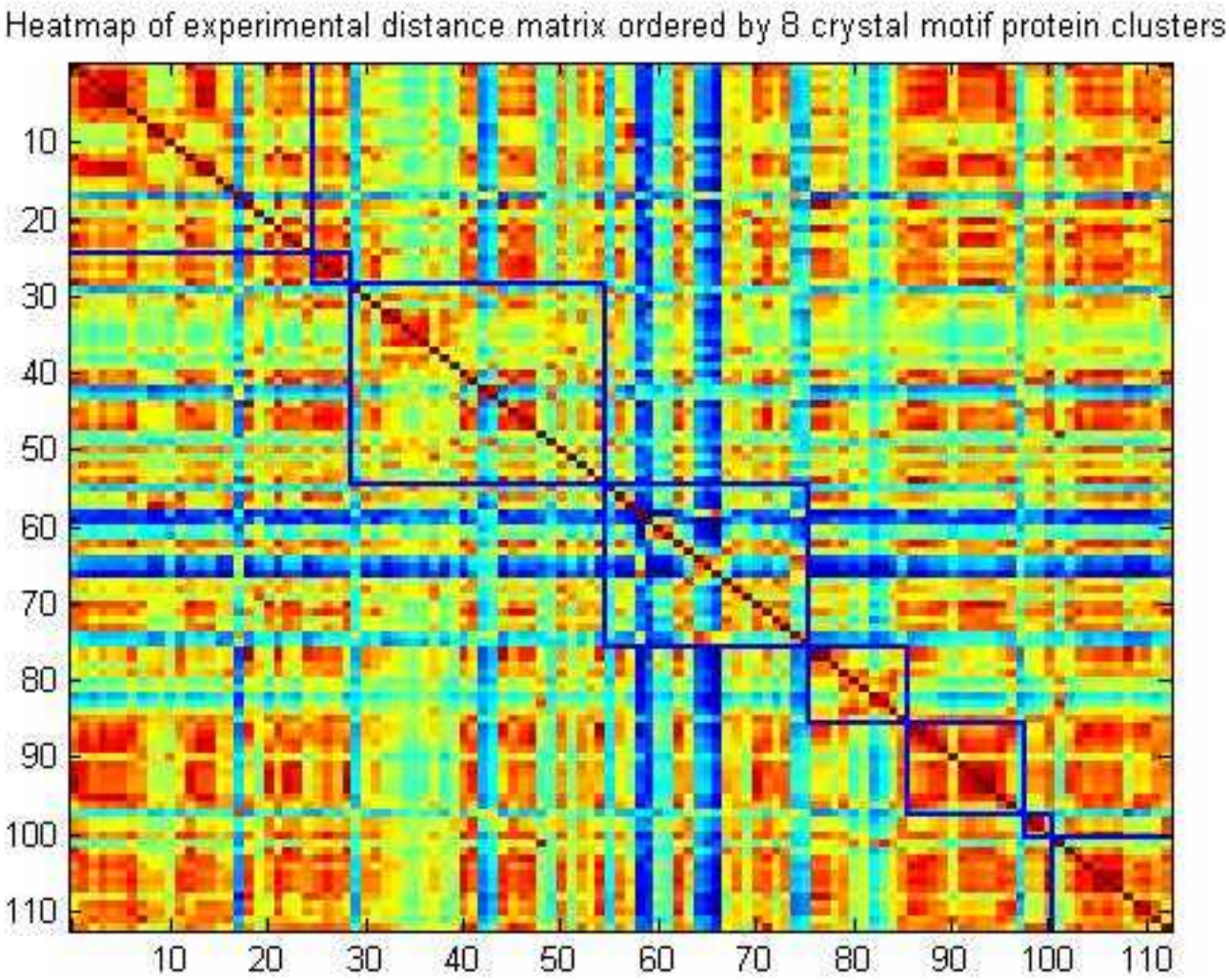
Kinase activity



KD Greis, et al., *J Am Soc Mass Spectrom* 2006, 17, 815–22
M Bantscheff, et al., *Nature Biotech,* 2007, 25, 1035-44

Clustering by Structure vs Binding Data

Clustering by structure does not capture experimental binding patterns



Coordination & Collaboration

- Please list internal or external collaborative efforts