REDESIGNING DRUG DESIGN

John D. Chodera
MSKCC Computational Biology Program
http://www.choderalab.org

Slides available at http://www.choderalab.org

DISCLOSURES:
- Scientific Advisory Board, Schrödinger
Sometimes, Drug Discovery Works Well

Bcr-Abl fusion constitutively activates ABL in CML patients, resulting in unchecked white blood cell proliferation.

Imatinib bound to c-Abl [PDB:1IEP]

Human kinome [518 kinases]

Imatinib [blockbuster drug]

Staurosporine [toxic natural product]

Nature Biotech 23:329, 2005
DRUG DISCOVERY USUALLY ENDS IN FAILURE

Total pharma R&D spending doubled to $65B over 2000-2010
FDA approvals of new molecular entities went down by half
Number of truly innovative new molecules remained constant at 5-6/year
2010-2015 has seen large reductions in pharma R&D in the US

Drugs per billion US$ R&D spending*

Number of drugs per billion US$ R&D spending*


0.1 1.0 10.0 100.0

FDA tightens regulation post-thalidomide
FDA clears backlog following PDUFA regulations plus small bolus of HIV drugs
First wave of biotechnology-derived therapies

DRUG DISCOVERY USUALLY ENDS IN FAILURE

ERROOM’S LAW

MOORE’S LAW
DRUG DISCOVERY USUALLY ENDS IN FAILURE

probability of stage success

Cost per launch (out of pocket) $24 $49 $146 $62 $128 $185 $235 $44 $873

DRUG DISCOVERY USUALLY ENDS IN FAILURE


Probability of stage success:
- Target-to-hit: 80%
- Hit-to-lead: 75%
- Lead optimization: 85%
- Preclinical: 69%
- Phase I: 54%
- Phase II: 34%
- Phase III: 70%
- Submission to launch: 91%

4.5 years and $219M

This stage has a 92% failure rate.

~2-4% overall success rate.
WE REGULARLY DESIGN PLANES, BRIDGES, AND BUILDINGS ON COMPUTERS
WE REGULARLY DESIGN PLANES, BRIDGES, AND BUILDINGS ON COMPUTERS

$10^3 - 10^6$ parts

WHY NOT SMALL MOLECULE DRUGS?

< 10^2 atoms
How can we bring drug design into the 21st century?
HOW CAN WE BRING DRUG DESIGN INTO THE 21ST CENTURY?
How can we design small molecules to have intended biological effects?
HOW CAN WE **DESIGN** SMALL MOLECULES TO HAVE INTENDED BIOLOGICAL EFFECTS?

Will it **bind** the target with high **affinity**?
How can we design small molecules to have intended biological effects?

Will it bind the target with high affinity?
Will its binding mode have the intended effect on the target?
HOW CAN WE DESIGN SMALL MOLECULES TO HAVE INTENDED BIOLOGICAL EFFECTS?

Will it bind the target with high affinity?
Will its binding mode have the intended effect on the target?
Does it produce the desired effect on cellular
How can we **design** small molecules to have intended biological effects?

Will it bind the target with high affinity?
Will its binding mode have the **intended effect** on the target?
Does it produce the **desired effect** on cellular?
Will it bind **unintended targets**? Are the resulting effects unacceptably toxic?
Multiscale physical models can drive small molecule design.

- Physical binding constant: $K_d$
- Catalytic life cycle: $K_{i,app}$
- Cellular pathways: EC50

We use physical modeling and statistical mechanics to build predictive models.
How can we compute binding affinities for molecules that have yet to be synthesized or tested?

Virtual screening methods are in widespread use in drug discovery efforts today. They must work well, right?

$P + L \xrightarrow{K_d} PL$
HOW CAN WE COMPUTE BINDING AFFINITIES FOR MOLECULES THAT HAVE YET TO BE SYNTHESIZED OR TESTED?

Virtual screening methods are in widespread use in drug discovery efforts today. They must work well, right?

\[ P + L \rightleftharpoons_{\text{rate}} K_d \text{PL} \]

**Figure 11.** Plot of scaled score vs pAffinity for MRS and PPAR\(\delta\). While the calculated correlation coefficient for the data shown for MRS is \(r = -0.28\), this plot clearly demonstrates that these values are meaningless. No useful correlation exists between the docking score and compound affinity.

“For prediction of compound affinity, none of the docking programs or scoring functions made a useful prediction of ligand binding affinity.”

How accurate does one need to be to have an impact on drug discovery?

A 2 kcal/mol error in prioritizing lead synthesis would speed lead optimization by 3x but even 10% improvements would be of tremendous benefit.

WHAT DETAILS ARE CRUCIAL FOR ACCURACY?

NATURE
- Born-Oppenheimer approximation
- Limited treatment of QM (DFT or semiempirical)
- Implicit treatment of some electrons

QM/MM
- Implicit representation of all electrons

POLARIZABLE MM
- Neglect of polarization
- Representation of multipolar moments by fixed charges

MM
- Rigid receptor
- Rigid or semi-rigid ligand
- Single-configuration scoring
- Simple solvation model

DOCKING
WHAT DETAILS ARE CRUCIAL FOR ACCURACY?

NATURE
- Born-Oppenheimer approximation
- Limited treatment of QM (DFT or semiempirical)
- Implicit treatment of some electrons

QM/MM
- Implicit representation of all electrons

POLARIZABLE MM
- Neglect of polarization
- Representation of multipolar moments by fixed charges

MM
- Rigid receptor
- Rigid or semi-rigid ligand
- Single-configuration scoring
- Simple solvation model

ENTROPY
- ENTHALPY
- CONFORMATIONAL HETEROGENEITY
WHAT DETAILS ARE CRUCIAL FOR ACCURACY?

- Born-Oppenheimer approximation
- Limited treatment of QM (DFT or semiempirical)
- Implicit treatment of some electrons
- Implicit representation of all electrons
- Neglect of polarization
- Representation of multipolar moments by fixed charges
- Rigid receptor
- Rigid or semi-rigid ligand
- Single-configuration scoring
- Simple solvation model

Molecular mechanics potential energy:

\[
V(q) = \sum_{\text{bonds}} K_r(r - r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[ \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right]
\]
How can we compute a binding affinity including relevant statistical mechanics?

\[ P + L \xrightleftharpoons[K_d]{\tau_{\text{unbound}}}[\tau_{\text{bound}}] PL \]

- **$K_d$**: dissociation constant

$$K_d \propto \frac{\tau_{\text{unbound}}}{\tau_{\text{bound}}}$$
How can we compute a binding affinity including relevant statistical mechanics?

For typical drug off-rates ($10^{-4}$ s$^{-1}$), reliable calculation of binding affinities would require hour trajectories, requiring $\sim 10^6$ years to simulate.
**ALCHEMICAL FREE ENERGY CALCULATIONS PROVIDE A RIGOROUS Way TO EFFICIENTLY COMPUTE BINDING AFFINITIES**

![Diagram](attachment:image.png)

\[ \Delta G_{bind} = P + L \rightarrow PL \]

\[ \Delta G_{bind} = P + \emptyset \rightarrow P\emptyset \]

**Requires orders of magnitude** less effort than simulating direct association process, but still includes all enthalpic/entropic contributions to binding free energy.

\[ \Delta F_{1\rightarrow N} = -\beta^{-1} \ln \frac{Z_N}{Z_1} = -\beta^{-1} \ln \frac{Z_2}{Z_1} \cdot \frac{Z_3}{Z_2} \cdots \frac{Z_N}{Z_{N-1}} = \sum_{n=1}^{N-1} \Delta F_{n\rightarrow n+1} \]

\[ Z_n = \int dx e^{-\beta U(x)} \]

Pioneering work from many: McCammon, van Gunsteren, Kollman, Jorgensen, Chipot, Roux, Boresch, Fujitani, Pande, Shirts, Swope, Christ, Mobley, and many more

**ALCHEMICAL FREE ENERGY CALCULATIONS PROVIDE A RIGOROUS WAY TO EFFICIENTLY COMPUTE BINDING AFFINITIES**

\[
\Delta G_{\text{bind}} = \Delta \left( P + L \rightarrow PL \right) = \Delta \left( P + \varnothing \rightarrow P\varnothing \right)
\]

multiple simulations of alchemical intermediates

restraint imposition

discharging

steric decoupling

noninteracting

Requires **orders of magnitude** less effort than simulating direct association process, but still includes all enthalpic/entropic contributions to binding free energy.

\[
\Delta F_{1\rightarrow N} = -\beta^{-1} \ln \frac{Z_N}{Z_1} = -\beta^{-1} \ln \frac{Z_2}{Z_1} \cdot \frac{Z_3}{Z_2} \cdot \ldots \cdot \frac{Z_N}{Z_{N-1}} = \sum_{n=1}^{N-1} \Delta F_{n\rightarrow n+1} = Z_n = \int dx \, e^{-\beta U(x)}
\]

Pioneering work from many: McCammon, van Gunsteren, Kollman, Jorgensen, Chipot, Roux, Boresch, Fujitani, Pande, Shirts, Swope, Christ, Mobley, and many more

ALCHEMICAL METHODS CAN ALSO COMPUTE MANY OTHER USEFUL PROPERTIES

partition coefficients (logP, logD) and permeabilities

selectivity for subtypes or related targets/off-targets

lead optimization of affinity and selectivity

susceptibility to resistance mutations

also solubilities, polymorphs, etc.
Alchemical free energy methods can work reliably in simple systems, but complex systems remain challenging.

Model systems

- hydration free energies
  - 1.04±0.03 kcal/mol (N=44)
    - Mobley et al. JPC B, 2007
  - 1.23±0.01 kcal/mol (N=502)
    - Mobley et al. JPC B 2009.

- T4 lysozyme L99A
  - 1.89±0.04 kcal/mol (N=13)

- FKBP12
  - 0.4 kcal/mol* (N=8)
    - Fujitani et al. JCP 123:084108, 2005
    - * with 3.2 kcal/mol offset

Pharmaceutically relevant

- JNK3 kinase
  - Anecdotal literature reports of success (publication bias?)

Retrospective RMS error [sample size] (not to scale)
**ALCHEMICAL FREE ENERGY METHODS CAN WORK RELIABLY IN SIMPLE SYSTEMS, BUT COMPLEX SYSTEMS REMAIN CHALLENGING**

**Model systems**

- **T4 lysozyme L99A**
  - $1.89 \pm 0.04$ kcal/mol ($N=13$)

- **FKBP12**
  - $0.4$ kcal/mol* ($N=8$)
  - Fujitani et al. JCP 123:084108, 2005
  - *with $3.2$ kcal/mol offset

- **JNK3 kinase**

**Pharmacologically relevant**

- **Hydration free energies**
  - $1.04 \pm 0.03$ kcal/mol ($N=44$)
    - Mobley et al. JPC B, 2007
  - $1.23 \pm 0.01$ kcal/mol ($N=502$)
    - Mobley et al. JPC B 2009.
  - $1.33 \pm 0.05$ kcal/mol ($N=17$)

**Future work**

- Anecdotal literature reports of success (publication bias?)

**Calculations are notoriously unreliable.**
(e.g. SAMPL predictive challenges)
Alchemical free energy methods can work reliably in simple systems, but complex systems remain challenging. Applications include:

- **Model systems**
  - Hydration free energies: solvent only, small, neutral molecules, fixed protonation states.
  - T4 lysozyme L99A: small, rigid protein, small, neutral ligands, fixed protonation states, multiple sidechain orientations, multiple ligand binding modes.
  - FKBP12: small, rigid protein, fixed protonation states, larger drug-like ligands, with few rotatable bonds.

- **Pharmaceutically relevant**
  - JNK3 kinase: large protein, multiple conformations, large drug-like ligands, rotatable bonds, multiple protonation states? tautomers? phosphorylation and activation, peptide substrate? MgCl$_2$ salt effects?

The diagrams illustrate the transition from simple to complex systems, highlighting the challenges and opportunities in each category.
There were 250 bridge failures in the US and Canada between 1878-1888.

“The subject of mechanical pathology is relatively as legitimate and important a study to the engineer as medical pathology is to the physician. While we expect the physician to be familiar with physiology, without pathology he would be of little use to his fellow-men, and it [is] as much within the province of the engineer to investigate causes, study symptoms, and find remedies for mechanical failures as it is to direct the sources of power in nature for the use and convenience of man.”

- George Thomson, 1888
“When I started Intel we couldn’t make a device twice in a row in the same way. I earned my reputation by being part of a team that figured out why a thing was not reproducible... The attitude [in high-tech] is, something went wrong for a reason, let’s find the gold nugget... But in pharma, if a clinical trial doesn’t work...they just throw [the drug] away...”

- Andy Grove (former CEO of Intel)
PREDICTIONS FAIL FOR THREE REASONS

1. The forcefield does a poor job of modeling the physics of our system

\[
V(q) = \sum_{\text{bonds}} K_r (r - r_{eq})^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[ \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right]
\]

2. We’re missing some essential chemical in our simulations (e.g. protonation states, tautomers, covalent association)

3. We haven’t sampled all of the relevant conformations

WE NEED TO UNDERSTAND WHY FAILURES OCCUR TO IMPROVE THE ROBUSTNESS OF OUR PREDICTIVE MODELS
FAIL FAST, FAIL CHEAP

computational predictions

experimental confirmation
HOW CAN WE SPEED UP FREE ENERGY CALCULATIONS?

$50K
5 TFLOP/S

DOUBLING
-18 MONTHS

many CPU-weeks/calculation

DOESN’T FIT NEATLY IN A SYNTHETIC CHEMIST’S TIMEFRAME TO WAIT WEEKS FOR AN ANSWER.
HOW CAN WE SPEED UP FREE ENERGY CALCULATIONS?

$50K
5 TFLOP/S

DOUBLING EVERY 18 MONTHS

many CPU-weeks/calculation

$500
5 TFLOP/S

BEATING MOORE’S LAW?

overnight on a workstation?

WE CAN EXPLOIT NEW GPU TECHNOLOGIES TO REACH PRACTICAL COMPUTATION TIMES
YANK: AN OPEN-SOURCE, COMMUNITY-ORIENTED PLATFORM FOR GPU-ACCELERATED FREE ENERGY CALCULATIONS

NVIDIA GTX-TITAN ($1000)

OpenMM speedup (GTX Titan) over 12-core Xeon X5650 CPU for DHFR

<table>
<thead>
<tr>
<th>method</th>
<th>natoms</th>
<th>gromacs CPU</th>
<th>OpenMM GPU</th>
<th>speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB/SA</td>
<td>2,489</td>
<td>2.54 ns/day</td>
<td>287 ns/day</td>
<td>113 x</td>
</tr>
<tr>
<td>RF</td>
<td>23,558</td>
<td>18.8 ns/day</td>
<td>163 ns/day</td>
<td>8.7 x</td>
</tr>
<tr>
<td>PME</td>
<td>23,558</td>
<td>6.96 ns/day</td>
<td>104 ns/day</td>
<td>15 x</td>
</tr>
</tbody>
</table>

http://openmm.org

A free, open-source, extensible platform for free energy calculations and ligand design

http://www.getyank.org
Hamiltonian exchange protocol allows for repeated binding/unbinding events and reorientation in site.

- **solid** fully interacting
- **transparent** noninteracting

Indole binding to T4 lysozyme L99A
12 h on 2 NVIDIA Tesla M2090 GPUs
Hamiltonian exchange with Gibbs sampling

Chodera and Shirts. JCP 135:194110, 2011
http://github.org/choderalab/yank
Hamiltonian exchange protocol allows for repeated binding/unbinding events and reorientation in site.

solid fully interacting
transparent noninteracting

indole binding to T4 lysozyme L99A
12 h on 2 NVIDIA Tesla M2090 GPUs
Hamiltonian exchange with Gibbs sampling

Chodera and Shirts. JCP 135:194110, 2011
http://github.org/choderalab/yank
Additional binding sites can be identified and individual affinities estimated by mixing in Monte Carlo moves.

Benzene bound to T4 lysozyme L99A
AMBER96 + OBC GBSA

Chodera and Shirts. JCP 135:194110, 2011
http://github.org/choderalab/yank
FREE ENERGIES WITH IMPLICIT MODELS OF SOLVENT ARE PROMISING: COULD PLAY A ROLE IN RAPID AFFINITY PREDICTION

T4 LYSOZYME L99A

FKBP12

AMBER ff96 + OBC GBSA (no cutoff) + GAFF/AM1-BCC
12 h on 2 GPUs

Chodera and Shirts. JCP 135:194110, 2011
http://github.org/choderalab/yank
FAIL FAST, FAIL CHEAP

computational predictions

experimental confirmation
1. PREDICT HOW MODIFICATIONS TO LIGAND WILL CHANGE AFFINITY
2. TEST EXPERIMENTALLY
1. PREDICT HOW MODIFICATIONS TO LIGAND WILL CHANGE AFFINITY
2. TEST EXPERIMENTALLY
SYNTHESIS OF NEW COMPOUNDS TO TEST HYPOTHESES IS EXPENSIVE AND TIME-CONSUMING

SYNTHESIS OF IMATINIB
1. PREDICT HOW MODIFICATIONS TO PROTEIN WILL CHANGE AFFINITY
2. TEST EXPERIMENTALLY
1. PREDICT HOW MODIFICATIONS TO **PROTEIN** WILL CHANGE AFFINITY
2. TEST EXPERIMENTALLY

\[ \Delta G_1 \]

\[ \Delta G_2 \] +

\[ \Delta G_3 \]
Inverting the drug discovery problem allows us to fail quickly and cheaply.

- Quick-change mutagenesis
- Expression assessment in 1 ml culture
- Expression and purification in ~2L culture
- Purchase known ligands

**Techniques:**

- Isothermal titration calorimetry
- Surface plasmon resonance
- Fluorescence binding assays
HOW CAN WE MAKE WETLAB EXPERIMENTS LOOK MORE LIKE PROBLEMS WE KNOW HOW TO SOLVE EFFICIENTLY?

messy
laborious
inconsistent
skill-dependent
9 am - 5 pm
HOW CAN WE MAKE WETLAB EXPERIMENTS LOOK MORE LIKE PROBLEMS WE KNOW HOW TO SOLVE EFFICIENTLY?

- messy
- laborious
- inconsistent
- skill-dependent
- 9 am - 5 pm

- precise
- structured
- consistent
- reproducible
- round-the-clock
Automated platform for bacterial cloning, mutagenesis, expression, purification, and binding affinity measurement with 24/7 operational capacity
Automated platform for bacterial cloning, mutagenesis, expression, purification, and binding affinity measurement with 24/7 operational capacity
ASSAY AUTOMATION CAN CONTROL ERROR
ASSAY AUTOMATION CAN CONTROL ERROR

Dispense  Shake/Mix  Spin  Read Fluorescence
MODEL SYSTEMS CAN TEACH US VALUABLE LESSONS

T4 LYSOZYME L99A
- small, rigid protein
- small, neutral ligands
- fixed protonation states
- multiple sidechain orientations
- multiple ligand binding modes

FKBP-12
- small, rigid protein
- fixed protonation states
- larger natural product-like ligands with rotatable bonds

IL-2
- small protein
- fixed protonation states
- some allostery and binding site plasticity

TRYPsin
- small, rigid protein
- small ligands
- charged ligands
- protonation state changes

KINASES
- large protein, multiple conformations
- large drug-like ligands, rotatable bonds
- multiple protonation states? tautomers?
- phosphorylation and activation peptide substrate?
WHERE DO MODEL SYSTEMS COME FROM?

- Word of mouth ("Hey, you should really look at aspartyl proteases...")
- My old advisor worked on this (T4 lysozyme mutants)
- I got the plasmid from the lab down the hall (chicken Src)
- Everybody else is working on it! (Abl)
WHERE DO MODEL SYSTEMS COME FROM?

• Word of mouth ("Hey, you should really look at aspartyl proteases...")
• My old advisor worked on this (T4 lysozyme mutants)
• I got the plasmid from the lab down the hall (chicken Src)
• Everybody else is working on it! (Abl)

SURELY THERE MUST BE A BETTER WAY!
CAN WE MINE PUBLIC DATASETS FOR GOOD MODEL SYSTEMS?

Desiderata:
• good bacterial expression (for cheap protein production)
• multiple structures available in PDB
• a variety of known ligands available for purchase
• large dynamic range of binding affinities (>3 kcal/mol)
• accessibility to biophysical assays (fluorescence, SPR, ITC)
• known point mutants (e.g. UniProt)
• disease relevance (for funding!)
• properties characteristic of real challenging targets
PANNING FOR MODEL SYSTEMS

- initial set of UniProt IDs
- retrieve all UniProt metadata
- retrieve all known structures and ligands
- retrieve additional data
- filter by criteria of interest
PANNING FOR MODEL SYSTEMS

- **PDB**
  - initial set of UniProt IDs
  - retrieve all UniProt metadata

- **UniProt**
  - retrieve all known structures and ligands

- **PDB**
  - expression/cofactors/etc

- **ChEMBL**
  - ZINC/eMolecules
  - retrieve additional data

- **filter**
  - filter by criteria of interest
SOME TARGETS HAVE BIOASSAY DATA FOR MULTIPLE FDA-APPROVED DRUGS
Many targets have usefully large dynamic ranges of known affinities
Many targets have usefully large dynamic ranges of known affinities
Can we search for potential fluorescent probe compounds?

Quinazoline scaffolds are often fluorescent...

...which can be expressed as a SMARTS query:

```
c1cccc2c1cncn2
```

...and used to find some quinazoline scaffold inhibitors of Uniprot P00918 (carbonic anhydrase II) to serve as probes:

567 nM

490 nM

700 nM

Can we expand this search to all known fluorescent scaffolds?

Thanks OpenEye!
MANY OF SYSTEMS CAN BE EXPRESSED BY ROBOTS USING A STANDARD PROTOCOL

<table>
<thead>
<tr>
<th>ID</th>
<th>Gene</th>
<th>Protein</th>
<th>PCR Result</th>
<th>Expression test</th>
<th>ug/mL culture</th>
<th>Tag (Best)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Abl1</td>
<td>Abl kinase</td>
<td>yes</td>
<td>NO</td>
<td>121</td>
<td>SUMO</td>
</tr>
<tr>
<td>2</td>
<td>AK1</td>
<td>adenylate kinase 1</td>
<td>yes</td>
<td>High</td>
<td>121</td>
<td>H6</td>
</tr>
<tr>
<td>3</td>
<td>AK2</td>
<td>adenylate kinase 2</td>
<td>yes</td>
<td>High</td>
<td>121</td>
<td>H6</td>
</tr>
<tr>
<td>4</td>
<td>AMPC_WT</td>
<td>AmpC beta-lactamase</td>
<td>NO</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>5</td>
<td>CAH2</td>
<td>carbonic anhydrase II</td>
<td>yes</td>
<td>High</td>
<td>198</td>
<td>H6</td>
</tr>
<tr>
<td>6</td>
<td>CALM1</td>
<td>calmodulin</td>
<td>yes</td>
<td>High</td>
<td>107</td>
<td>SUMO</td>
</tr>
<tr>
<td>7</td>
<td>CCP_WT</td>
<td>cytochrome c peroxidase</td>
<td>yes</td>
<td>High</td>
<td>224</td>
<td>H6</td>
</tr>
<tr>
<td>8</td>
<td>CCP_GA</td>
<td>cytochrome c peroxidase</td>
<td>yes</td>
<td>High</td>
<td>198</td>
<td>H6</td>
</tr>
<tr>
<td>9</td>
<td>CDK2</td>
<td>cyclin-dependent kinase 2</td>
<td>yes</td>
<td>Low</td>
<td>19</td>
<td>SUMO</td>
</tr>
<tr>
<td>10</td>
<td>DHFR</td>
<td>dihydrofolate reductase</td>
<td>yes</td>
<td>Low</td>
<td>5</td>
<td>MOCR</td>
</tr>
<tr>
<td>11</td>
<td>ESR1</td>
<td>estrogen receptor α</td>
<td>yes</td>
<td>NO</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>12</td>
<td>ESR2</td>
<td>estrogen receptor β</td>
<td>yes</td>
<td>NO</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>13</td>
<td>FKBP12</td>
<td>FK506 binding protein</td>
<td>WRONG ORF</td>
<td>WRONG ORF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>GYRB</td>
<td>E. coli DNA gyrase B</td>
<td>yes</td>
<td>High</td>
<td>152</td>
<td>MOCR</td>
</tr>
<tr>
<td>15</td>
<td>HSV1_TK</td>
<td>thymidine kinase</td>
<td>yes</td>
<td>High</td>
<td>161</td>
<td>MOCR</td>
</tr>
<tr>
<td>16</td>
<td>IL2</td>
<td>interleukin 2</td>
<td>yes</td>
<td>Low-Moderate</td>
<td>29</td>
<td>MOCR</td>
</tr>
<tr>
<td>17</td>
<td>JNK3</td>
<td>JNK3 kinase</td>
<td>yes</td>
<td>Low</td>
<td>9</td>
<td>MOCR</td>
</tr>
<tr>
<td>18</td>
<td>p38MAPK14</td>
<td>p38 kinase</td>
<td>yes</td>
<td>High</td>
<td>170</td>
<td>MOCR</td>
</tr>
<tr>
<td>19</td>
<td>PIM1</td>
<td>PIM1 kinase</td>
<td>yes</td>
<td>High</td>
<td>102</td>
<td>MOCR</td>
</tr>
<tr>
<td>20</td>
<td>RXRa</td>
<td>retinoic acid receptor</td>
<td>yes</td>
<td>High</td>
<td>88</td>
<td>TRX</td>
</tr>
<tr>
<td>21</td>
<td>SRC</td>
<td>Src kinase</td>
<td>NO</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>22</td>
<td>T4_L99A</td>
<td>T4 lysozyme L99A</td>
<td>yes</td>
<td>Low-Moderate</td>
<td>42</td>
<td>H6</td>
</tr>
<tr>
<td>23</td>
<td>T4_L99A/M102Q</td>
<td>T4 lysozyme L99A/M102Q</td>
<td>yes</td>
<td>Low-Moderate</td>
<td>18</td>
<td>H6</td>
</tr>
<tr>
<td>24</td>
<td>TRYP3 (PRSS3)</td>
<td>trypsin</td>
<td>yes</td>
<td>NO- Very Low</td>
<td></td>
<td>TRX</td>
</tr>
</tbody>
</table>
WHAT MATTERS?

NEED SENSITIVITY ANALYSIS

* KINASE CONFORMATION
* PROTONATION STATES
* KINASE
* INHIBITOR
* SALT ENVIRONMENT
* SOLVENT MODEL
* ELECTROSTATIC TREATMENT
* FORCEFIELD
* KINASE PHOSPHORYLATION STATE

Afinitin (EGFR, 2013)
Ibrutinib (BTK, 2013)
Pazopanib (VEGFR, 2009)

Afatinib (EGFR, 2013)
Idelalisib (PI3K, 2014)

Bosutinib (ABL, 2012)
Imatinib (ABL, 2001)

Regorafenib (VEGFR, 2012)

Cabozaantinib (MET, 2012)
Lapatinib (HER2, 2007)

Ruxolitinib (JAK1/2, 2011)

Ceritinib (ALK, 2014)

Sorafenib (VEGFR, 2005)

Crizotinib (ALK, MET, 2011)

Sunitinib (VEGFR, 2006)

Dabrafenib (BRAF, 2013)

Tofacitinib (JAK3, 2012)

Dasatinib (ABL, 2006)

Trametinib (MEK, 2013)

Erlotinib (EGFR, 2005)

Nilotinib (ABL, 2007)

Vandetanib (VEGFR, 2011)

Gefitinib (EGFR, 2003)
Palbociclib (CDK4/6, 2015)

Lenvatinib (VEGFR, 2015)

Nilotinib (ABL, 2007)

Vemurafenib (BRAF, 2011)

Regorafenib (VEGFR, 2012)

Trametinib (MEK, 2013)

Vandetanib (VEGFR, 2011)

Vemurafenib (BRAF, 2011)

Regorafenib (VEGFR, 2012)

Trametinib (MEK, 2013)

Vandetanib (VEGFR, 2011)

Vemurafenib (BRAF, 2011)
PREDICTIONS FAIL FOR THREE REASONS

1. The forcefield does a poor job of modeling the physics of our system

\[ V(q) = \sum_{\text{bonds}} K_r (r - r_{eq})^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[ \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right] \]

2. We’re missing some essential chemical in our simulations (e.g. protonation states, tautomers, covalent association)

3. We haven’t sampled all of the relevant conformations
PREDICTIONS FAIL FOR THREE REASONS

1. The forcefield does a poor job of modeling the physics of our system

\[
V(q) = \sum_{\text{bonds}} K_r (r - r_{eq})^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)] + \sum_{i < j} \left[ \frac{A_{ij}}{R_{ij}^1} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right]
\]
HOW ARE FORCEFIELDS MADE?

- experimental data
- quantum chemistry
- keen chemical intuition
- heroic effort by graduate students and postdocs

a parameter set we desperately hope someone actually uses
AS DRUG DISCOVERY EXPLORES NEW PARTS OF CHEMICAL SPACE, HOW CAN FORCEFIELDS KEEP UP?

The Generalized Amber Forcefield (GAFF) was parameterized with this chemical universe:

Extension of this universe is nontrivial because parameter fitting code never released!

The approach to parameterization has evolved over time, but it’s still not completely automated by any measure.

<table>
<thead>
<tr>
<th>Year</th>
<th>Forcefield</th>
<th>Parameter Fitting</th>
<th>Atom Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990s</td>
<td>AMBER parm96</td>
<td>lots of “hand tweaking”</td>
<td>hand-picked</td>
</tr>
<tr>
<td>Early 2000s</td>
<td>GAFF</td>
<td>genetic algorithm</td>
<td>hand-picked</td>
</tr>
<tr>
<td>Mid 2000s</td>
<td>TIP4P-Ew</td>
<td>least-squares optimization</td>
<td>hand-picked</td>
</tr>
</tbody>
</table>

How can we move to automated schemes that are easy to grow and refine?
THE APPROACH TO PARAMETERIZATION HAS EVOLVED OVER TIME, BUT IT’S STILL NOT COMPLETELY AUTOMATED BY ANY MEASURE

<table>
<thead>
<tr>
<th>year</th>
<th>forcefield</th>
<th>parameter fitting</th>
<th>atom types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990s</td>
<td>AMBER parm96</td>
<td>lots of “hand tweaking”</td>
<td>hand-picked</td>
</tr>
<tr>
<td>early 2000s</td>
<td>GAFF</td>
<td>genetic algorithm</td>
<td>hand-picked</td>
</tr>
<tr>
<td>mid 2000s</td>
<td>TIP4P-Ew</td>
<td>least-squares optimization</td>
<td>hand-picked</td>
</tr>
</tbody>
</table>

Torsion barrier for peptide bond from parm96.dat

| X | -C | -N | -X | 4  | 10.00 | 180.0 | 2. | AA | check Wendy? & NMA |

How can we move to automated schemes that are easy to grow and refine?
WHAT DO WE WANT OUT OF A FORCEFIELD PARAMETERIZATION SCHEME?

Everything is automatic; don’t need to tweak things by hand.

Stupendous feats of chemical insight are not required.

Automatically chooses optimal functional forms.

We can add more data when we reach uncharted parts of chemical space.

Would give us an idea of how reliable new predictions are expected to be.

We can build a map of what data we should try to collect to improve accuracy.

Is there a procedure that could fit these criteria?
THE OLD WAY

\[ \theta^* \rightarrow \text{MD} \rightarrow \text{computed result} \rightarrow \langle A \rangle_\star \]

One set of parameters in, one computed result out
Bayes rule provides a **probability measure** over unknown parameters given data and an automated way to **update** parameters given new experimental data.

\[ p(\theta|D) \propto p(D|\theta)p(\theta) \]
THE BAYESIAN WAY

\[ \begin{align*}
\theta_1 & \rightarrow \langle A \rangle_1 \\
\theta_2 & \rightarrow \langle A \rangle_2 \\
\theta_3 & \rightarrow \langle A \rangle_3 \\
\theta_4 & \rightarrow \langle A \rangle_4
\end{align*} \]

Multiple parameter sets in, multiple estimates out

We can estimate both statistical and systematic components of computed results.
WHERE DO WE GET THE DATA?
Where do we get the data?

“ANALOGUE DATABASES”
The literature is filled with erroneous data.
DATA HAS A HABIT OF BEING RE/MISREPORTED: 
THE GENEALOGY OF A SINGLE MEASUREMENT

EXPLANATION

Presumed
Superscripts:
a-DDE $K_{ow}$ value only
b-DDT $K_{ow}$ value only
c-DDT and DDE $K_{ow}$ values

NIST HAS A SOLUTION

with Kenneth Kroenlein, NIST TRC

JPC B ASAP. DOI:10.1021/acs.jpcb.5b06703

<table>
<thead>
<tr>
<th>Filter step</th>
<th>Mass density</th>
<th>Static dielectric</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Single Component</td>
<td>136212</td>
<td>1651</td>
</tr>
<tr>
<td>2. Druglike Elements</td>
<td>125953</td>
<td>1651</td>
</tr>
<tr>
<td>3. Heavy Atoms</td>
<td>71595</td>
<td>1569</td>
</tr>
<tr>
<td>4. Temperature</td>
<td>38821</td>
<td>964</td>
</tr>
<tr>
<td>5. Pressure</td>
<td>14103</td>
<td>461</td>
</tr>
<tr>
<td>6. Liquid state</td>
<td>14033</td>
<td>461</td>
</tr>
<tr>
<td>7. Aggregate T, P</td>
<td>3592</td>
<td>432</td>
</tr>
<tr>
<td>8. Density+Dielectric</td>
<td>246</td>
<td>246</td>
</tr>
</tbody>
</table>
DENSITIES OF MOLECULAR LIQUIDS ARE REASONABLY WELL MODELED

![Graph showing comparison between experimental and predicted densities of molecular liquids]

Experiment (ThermoML) vs. Predicted (GAFF)

Density [g/cm³] (relative rms: 0.030 ± 0.001)

JPC B ASAP. DOI:10.1021/acs.jpcb.5b06703
LOW-DIELECTRIC MOLECULES ARE POORLY MODELED

![Inverse Static Dielectric Constant](chart.png)

\[ U(r) = \frac{q_1 q_2}{\varepsilon r} \propto \frac{1}{\varepsilon} \]

JPC B ASAP. DOI:10.1021/acs.jpcb.5b06703
NEW DATA WILL GREATLY IMPROVE FORCEFIELD QUALITY

Temperature-dependent densities of binary mixtures provides valuable information about atomic interactions

Patrick Grinaway

Julie Behr

Mettler-Toledo DM40 density meter
- accuracy: 0.0001 g/cm³
- range: 0-3 g/cm³
- temperature: 0-91 °C
- sample volume: 1 mL

Mettler-Toledo SC30
automated 30-sample changer
THE FUTURE OF FORCEFIELD PARAMETERIZATION?

EXPERIMENTAL DATA
QUANTUM CHEMISTRY
UNCERTAINTIES

ENSEMBLE OF
PARAMETER SETS

BAYESIAN INFEERENCE;
CONTINUOUS AUTOMATIC UPDATING

PULL THE TRIGGER AND GO!
PREDICTIONS FAIL FOR THREE REASONS

1. The **forcefield** does a poor job of modeling the physics of our system

\[
V(q) = \sum_{\text{bonds}} K_r(r - r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[ \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right]
\]

2. We’re missing some **essential chemical** in our simulations (e.g. protonation states, tautomers, covalent association)

3. We haven’t **sampled** all of the relevant conformations
2. We’re missing some essential chemical in our simulations (e.g. protonation states, tautomers, covalent association)
Protonation state effects can be important for ligands.
PROTONATION STATE EFFECTS CAN BE IMPORTANT FOR LIGANDS
PROTONATION STATE EFFECTS CAN BE IMPORTANT FOR LIGANDS

(A) Imatinib

Inhibitor

Kinase

Complex

\[ L \rightleftharpoons HL^+ \rightleftharpoons \cdots \]

\[ \text{Abl}^{(DFG-in; Asp^N)} \rightleftharpoons \text{Abl}^{(DFG-out; Asp^N)} \]

\[ \text{Abl}^{(DFG-in; Asp^-)} \rightleftharpoons \text{Abl}^{(DFG-out; Asp^-)} \]

\[ HL^+:\text{Abl}^{(DFG-out; Asp^N)} \rightleftharpoons HL^+:\text{Abl}^{(DFG-out; Asp^-)} \rightleftharpoons \cdots \]
**LET'S NOT FORGET TAUTOMERS**

The tautomers of warfarin show the variety of possible structures.

**Fig. 13** The frequency distribution of tautomers of marketed drugs

**MORE THAN HALF OF ALL DRUGS HAVE 2 OR MORE TAUTOMERS**
Markov chain Monte Carlo (MCMC) provides a flexible framework for enhancements. Generalized hybrid Monte Carlo (GHMC) and ligand MC with ligand displacement/rotation allow sampling different binding sites. Monte Carlo protein and ligand titration attempts speed up sampling. Monte Carlo ligand tautomer change attempts include chemical effects. We can use free energy calculations and experiments to quantify the error in neglecting of protomers and tautomers.
PREDICTIONS FAIL FOR THREE REASONS

1. The forcefield does a poor job of modeling the physics of our system

\[ V(q) = \sum_{\text{bonds}} K_r (r - r_{eq})^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[ \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right] \]

2. We’re missing some essential chemical in our simulations (e.g. protonation states, tautomers, covalent association)

3. We haven’t sampled all of the relevant conformations
PREDICTIONS FAIL FOR THREE REASONS

3. We haven’t sampled all of the relevant conformations
HOW CAN WE QUANTITATIVELY UNDERSTAND (AND DESIGN) THE SELECTIVITY OF KINASE INHIBITORS?
INTERESTINGLY, SELECTIVITY BY THIS MEASURE VARIED MORE THAN FOLD AMONG THESE SEVEN SMALL MOLECULE KINASE INHIBITORSTESTED HERETHAT ARE CURRENTLY APPROVED FOR USE IN HUMANS. SUPPLEMENTARY ONLINE FROM FOR LAPATINIB TO FOR SUNITINIB WITH A FAIRLY EVEN DISTRIBUTION OVER THIS RANGE. THEN ONLY HIGH AFFINITY INTERACTIONS WERE TAKEN INTO ACCOUNT AND SELECTIVITY SCORES WERE CALCULATED FOR BINDING INTERACTIONS WITH +D+N-IT. IT BECAME APPARENT THAT MOST OF THE COMPOUNDS BOUND ONLY ARE RELATIVELY SMALL NUMBER OF KINASES WITH HIGH AFFINITY. AMONG THE EXCEPTIONS WERE SUNITINIB AND DASATINIB. PRYCEL WAS ONE OF THE MARKETED DRUGS WHICH BOTH BOUND OF KINASES TESTED WITH +D+N- AND RESPECTIVELY. WHEREAS VALUES FOR GEFITINIB)RESSA (N+2-0- N+2-0- (

GATURE 3MALLMOLECULEnKINASEINTERACTIONMAPSFORKINASEINHIBITORS+INASESFOUNDTOBINDAREMARKEDWITHREDCIRCLES WHERELARGERCIRCLESINDICATEHIGHERAFFINITYBINDING)NTERACTIONSWITH#OMPLETERESULTSCANBEFOUNDIN#UPPLEMENTARY4ABLEONLINE4HEDATASETISALSOAVAILABLETHROUGHANINTERACTIVEWEBSITE HTTPWWWAMBITBIOCOMTECHNOLOGYPUBLICATIONS4HEKINASEDENDROGRAMWASADAPTEDANDISREPRODUCEDWITHPERMISSIONFROMSCIENCETHROUGHANINTERACTIVEWEBSITE HTTPWWWSCIENCEMAGORGAND#ELL3IGNALING4ECHNOLOGY INC HTTPWWWCELLSIGNALCOM

<table>
<thead>
<tr>
<th>ABT-869</th>
<th>AMG-706</th>
<th>AST-487</th>
<th>AZD-11529QPA</th>
<th>BIRB-796</th>
<th>BMS-597032</th>
<th>SHH-003</th>
<th>CHIR-256/TKI-258</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHIR-265/RAF-265</td>
<td>CI-1033</td>
<td>CP-690650</td>
<td>CP-724714</td>
<td>Dasatinib</td>
<td>EKB-569</td>
<td>Erlotinib</td>
<td></td>
</tr>
<tr>
<td>Erlotinib</td>
<td>Ekb-569</td>
<td>Fizzledol</td>
<td>Gatifinb</td>
<td>GW-2580</td>
<td>GW-788034</td>
<td>Imapinib</td>
<td>JNJ-7706621</td>
</tr>
<tr>
<td>JNJ-7706621</td>
<td>Lapatinib</td>
<td>LY-333531</td>
<td>MLN-518</td>
<td>MLN-8054</td>
<td>PI-103</td>
<td>PKC-412</td>
<td>PTK-787</td>
</tr>
<tr>
<td>PTK-787</td>
<td>Roscovitine/CYC-202</td>
<td>SB-202190</td>
<td>SB-203580</td>
<td>SB-431542</td>
<td>Sorafinib</td>
<td>Staurosporine</td>
<td>SU-14813</td>
</tr>
<tr>
<td>SU-14813</td>
<td>SU-14813</td>
<td>VX-745</td>
<td>VX-745</td>
<td>VX-680/MK-0457</td>
<td>ZD-6474</td>
<td>ZD-6474</td>
<td></td>
</tr>
</tbody>
</table>

$K_i$

1 nM
10 nM
100 nM
1 μM
10 μM
Differences in stabilities of inactive states may be responsible for origin of some kinase inhibitor selectivity. Essentially same binding mode in X-ray structure, same interactions. Calculations suggest no difference in binding free energy [J Biol Chem 285:13807, 2010; PNAS 110:1664, 2013].

\[
\begin{align*}
\text{Abl: imatinib} & \quad \Delta G = -10.9 \text{ kcal/mol} \\
\text{Src: imatinib} & \quad \Delta G = -6.2 \text{ kcal/mol} \\
\Delta \Delta G & \quad 4.7 \text{ kcal/mol}
\end{align*}
\]
Differences in stabilities of inactive states may be responsible for origin of some kinase inhibitor selectivity.

\[ \Delta G \text{ of confinement to binding-competent state} \]

\[ \Delta G \text{ of binding to binding-competent state} \]

\[ \text{net } \Delta G \text{ of binding} \]

Quantifying conformation energetics may be crucial to successful design of selective kinase inhibitors.

CAN WE BUILD AN ATLAS OF KINASE STRUCTURES AND ENERGETICS?

DIVIDE-AND-CONQUER TO COMPUTE AFFINITIES AND SELECTIVITIES

DIVIDE-AND-CONQUER TO COMPUTE AFFINITIES AND SELECTIVITIES

\[ \Delta G = -k_B T \ln \sum_i e^{-\beta (\Delta G_{i^{\text{conf}}} + \Delta G_{i^{\text{bind}}})} \]

**ΔG of trapping conformation**

**ΔG of binding to trapped conformation**

STRUCTURAL DATA ON HUMAN KINASES IS INCOMPLETE
STRUCTURAL DATA ON HUMAN KINASES IS INCOMPLETE

http://www.sgc.ox.ac.uk/research/kinases/
ENSEMBLER: AUTOMATING SIMULATIONS AT THE SUPERFAMILY SCALE

Patrick Grinaway
PBSB student

Daniel Parton
Postdoc

Kyle Beauchamp
Postdoc

Sonya Hanson
Postdoc

http://biorxiv.org/content/early/2015/06/29/018036
http://github.com/choderalab/ensembler
ENSEMBLER: AUTOMATING SIMULATIONS AT THE SUPERFAMILY SCALE

Patrick Grinaway
PBSB student

Daniel Parton
Postdoc

Patrick Gi
PBSB stu

Kyle Bea
Postdoc

Sonya Har
Postdoc

http://biornxiv.org/content/early/2015/06/29/018036
http://github.com/choderalab/ensembler
FOLDING@HOME GIVES US ACCESS TO ENORMOUS COMPUTATIONAL RESOURCES FOR PROBING BIOMOLECULAR DYNAMICS

Vijay S. Pande
Stanford University

Table last updated at Mon, 01 Jun 2015 23:02:21

OVER 31 PFLOP/S OF AGGREGATE COMPUTATIONAL POWER!

http://folding.stanford.edu
FOLDING@HOME ENABLES WHOLE-KINOME SIMULATION

518 human protein kinases excluding splice and disease variants

× 3,507 kinase catalytic domain structures in UniProt

= 1,816,626 kinase models will be built and refined on new MSKCC compute resources housed at SDSC

~ 18,166,260 kinase simulations on Folding@Home over one year
Exploiting scalable, fault-tolerant frameworks (e.g. hadoop, spark, redis, celery, cassandra) is essential to enable scalability to the family scale.
AN ADAPTIVE APPROACH CAN BE USED TO REDUCE STATISTICAL DYNAMICS TO A DISCRETE-STATE STOCHASTIC (MARKOV) MODEL

Similar in spirit to adaptive mesh refinement algorithms in engineering, very different from traditional approach of running a single long simulation.

with important contributions from Schütte, Noé, Weber, Hummer, Roux, Vanden-Eijnden

Many distinct metastable states can be identified at \( T \approx T_m \).

DIVIDE-AND-CONQUER
TO COMPUTE AFFINITIES AND SELECTIVITIES

DIVIDE-AND-CONQUER
TO COMPUTE AFFINITIES AND SELECTIVITIES

\[ \Delta G = -k_B T \ln \sum_i e^{-\beta (\Delta G_i^{\text{conf}} + \Delta G_i^{\text{bind}})} \]
CAN WE DRUG THE UNDRUGGABLE? ALLOSTERIC MODULATORS OF K-RAS MAY OPEN NEW DOORS IN CANCER THERAPY

Patrick Grinaway
In collaboration with Jeremy C. Smith (ORNL), Guillermo Perez-Hernandez and Frank Noé (FU Berlin)

human HRAS with GTP analogue [121p]

ORNL TITAN: 18,688 NVIDIA TESLA K20 GPUs
IS THERE A BETTER WAY?
HOW CAN WE **DESIGN** MOLECULES WITH DESIRED AFFINITIES AND SELECTIVITIES?

alternative substituents

alternative starting materials

How can we search large chemical spaces based on free energy objectives?
HOW CAN WE **DESIGN** MOLECULES WITH DESIRED AFFINITIES AND SELECTIVITIES?

Can we sample the joint space of configuration $x$ and chemical state $i$ so that the marginal chain maximizes a desired objective function?

$$\pi(i) \propto \int dx \pi(x, i)$$

**Express objectives in terms of ratios of partition functions:**

Maximize target affinity

$$\pi(i) \propto \frac{Z_{PL(i)}}{Z_L(i)} \propto K_a^{(i)}$$

Maximize selectivity for target protein (or conformation) $P_1$ over antitarget $P_2$

$$\pi(i) \propto \frac{Z_{P1L(i)}}{Z_{P2L(i)}} \propto \frac{K_a^{(i)}}{K_a^{(i)}}$$

Select resistance mutations that minimize inhibitor affinity while maintaining activity

$$\pi(i) \propto \frac{Z_{P(i)S}}{Z_{P(i)I}} \propto \frac{K_S^{(i)}}{K_i^{(i)}}$$
HOW CAN WE **DESIGN** MOLECULES WITH DESIRED AFFINITIES AND SELECTIVITIES?

Self-adjusted mixture sampling allows us to construct a recursion scheme to achieve a desired marginal distribution in terms of ratios of partition functions:

**Sample a new configuration with MCMC (e.g. hybrid Monte Carlo)**

\[ x_{n+1} \sim p(x|s_n) \]

**Sample a new chemical state with Monte Carlo**

\[ s_{n+1} \sim p(s|x_{n+1}, \{Z_s\}, \{\pi_s\}) \propto \frac{\pi_s}{Z_s} e^{-u_s(x)} \]

**Update free energy estimates using recursion**

\[ \log Z_{s,n+1} = \log Z_{s,n} - \frac{1}{n} \frac{w_s}{\pi_s} \]

**Update sampling target probabilities to maximize objective using recursion**

\[ \pi_{s,n+1} = \frac{Z_{PL}^{(s)}}{Z_L^{(s)}} \]

---

SAMPLING A NEW CHEMICAL STATE WITH REVERSIBLE JUMP MONTE CARLO

Abl kinase
imatinib >> nilotinib
Abl kinase
imatinib >> nilotinib
NONEQUILIBRIUM CANDIDATE MONTE CARLO CAN EXPONENTIALLY INCREASE ACCEPTANCE PROBABILITIES

Acceptance probability can be increased from $10^{-27}$ to $10^0$ (38%)!

Full 3D system
216 WCA particles
Reduced density
Reduced temperature
5 kT barrier
Nonequilibrium candidate Monte Carlo is an efficient tool for equilibrium simulation.
Treatment of CML with imatinib often induces resistance, predominantly E255K, T315I. Second-line drugs elicit further resistance:

**CAN WE DEVELOP A PHYSICAL MODEL OF RESISTANCE?**

We can hypothesize and test a simple physical mechanism of resistance: Resistance mutations reduce inhibitor binding affinity but retain ATP affinity (a surrogate for activity)

* Are certain inhibitors or binding modes less likely to elicit resistance?
* Can we incorporate likelihood of eliciting resistance mutations into rational ligand design?
CAN WE DEVELOP A PHYSICAL MODEL OF RESISTANCE?

We can hypothesize and test a simple physical mechanism of resistance:
Resistance mutations reduce inhibitor binding affinity but retain ATP affinity (a surrogate for activity)
* Are certain inhibitors or binding modes less likely to elicit resistance?
* Can we incorporate likelihood of eliciting resistance mutations into rational ligand design?
HIGH-THROUGHPUT FLUORESCENCE ASSAYS CAN MEASURE BINDING AFFINITIES TO MUTANT KINASES

fluorescent inhibitors

erlotinib

gemcitabine

gemcitabine

gemcitabine

gemcitabine

gemcitabine

gemcitabine
HIGH-THROUGHPUT FLUORESCENCE ASSAYS CAN MEASURE BINDING AFFINITIES TO MUTANT KINASES

excitation at 280 nm to achieve FRET from tyr/trp residues

https://github.com/choderalab/assaytools
High-throughput fluorescence assays can measure binding affinities to mutant kinases.
Bayesian inference allows us to quantify experimental uncertainty.

```python
# Sample with MCMC
mcmc = pymc.MCMC(pymc_model, db='ram', name='Sampler', verbose=True)
mcmc.sample(iter=10000, burn=10000, thin=50, progress_bar=False)
```

https://github.com/choderalab/assaytools
ASSAYS ARE EXPENSIVE: CAN WE CHANGE THAT?

assay plate: $10/plate
protein cost: $22/plate

https://github.com/choderalab/paper-microfluidics
ASSAYS ARE EXPENSIVE: CAN WE CHANGE THAT?

assay plate: $10/plate
protein cost: $22/plate

Paper Microzone Plates
Emanuel Carrilho, Scott T. Phillips, Sarah J. Vella, Andres W. Martinez, and George M. Whitesides

Anal. Chem. 2009, 81, 5990–5998

https://github.com/choderalab/paper-microfluidics
NEW TECHNOLOGIES FACILITATE RAPID PROGRESS

https://github.com/choderalab/paper-microfluidics
NEW TECHNOLOGIES FACILITATE RAPID PROGRESS

https://github.com/choderalab/paper-microfluidics
ASSAYS ARE EXPENSIVE: CAN WE CHANGE THAT?

Xerox ColorQube 8570 wax printer → patterned wax microplate → reflow wax through paper with hot plate

read in plate reader

Tecan HP D300 inkjet printer to dispense compounds in DMSO

https://github.com/choderalab/paper-microfluidics
ASSAYS ARE EXPENSIVE: CAN WE CHANGE THAT?

assay plate: $10/plate
protein cost: $22/plate

paper microzone plate: $0.20/plate
protein cost: $2/plate
10x reduction in costs!

https://github.com/choderalab/paper-microfluidics
Open source, high performance, high usability toolkits for predictive biomolecular simulation.

http://omnia.md
COLLABORATORS

http://folding.stanford.edu

Stanford
Vijay Pande
Sergio Bacallado
NIH SimBios

IBM Almaden
Bill Swope
Jed Pitera
Julia Rice

University of Chicago
Nina Singhal Hinrichs

UC Irvine
David Mobley

CCNY
Marilyn Gunner

OpenEye
Christopher Bayly
Anthony Nicholls

Stony Brook
Ken Dill
Markus Seeliger

UCSF
Brian Shoichet
David Sivak

University of Virginia
Michael Shirts

Duke
David Minh

Freie Universität Berlin
Frank Noé
Bettina Keller
Jan-Hendrik Prinz

Rutgers
Zhiqiang Tan

University of Edinburg
Antonia S. J. S. Mey

UC Berkeley
Phillip Elms (BioRad)
Susan Marqusee
Carlos Bustamante
Christian Kaiser
Gheorghe Christol

University of Chicago
Suri Vaikuntanathan

LBNL
Gavin Crooks

Vanderbilt
Joel Tellinghuisen

Hessian Informatics
Kim Branson

Vertex Pharmaceuticals
Richard Dixon

Code and data available at http://www.choderalab.org
WHICH KINASES CAN WE EASILY EXPRESS IN BACTERIA?
High yield bacterial expression of active c-Abl and c-Src tyrosine kinases

MARKUS A. SEELIGER,1,2 MATTHEW YOUNG,1,2 M. NIDANIE HENDERSON,1,2 PATRICIA PELLICENA,1,2 DAVID S. KING,1 ARNOLD M. FALICK,1,2 JOHN KURIYAN1,2,3

1 Howard Hughes Medical Institute, and 2 Department of Molecular and Cell Biology and Department of Chemistry, University of California, Berkeley, California 94720, USA
3 Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

(RECEIVED August 3, 2005; FINAL REVISION September 2, 2005; ACCEPTED September 4, 2005)

Abstract

The Abl and Src tyrosine kinases are key signaling proteins that are of considerable interest as drug targets in cancer and many other diseases. The regulatory mechanisms that control the activity of these protein kinases are complex, and involve large-scale conformational changes in response to phosphorylation and other modulatory signals. The success of the Abl inhibitor imatinib in the treatment of chronic myelogenous leukemia has shown the potential of kinase inhibitors, but the rise of drug resistance in patients has also shown that drugs with alternative modes of binding to the kinase are needed. The detailed understanding of the mechanisms of protein–drug interaction and drug resistance through biophysical methods demands a method for the production of active protein on the milligram scale. We have developed a bacterial expression system for the kinase domains of c-Abl and c-Src, which allows for the quick expression and purification of active wild-type and mutant kinase domains by coexpression with the YopH tyrosine phosphatase. This method makes practical the use of isotopic labeling of c-Abl and c-Src for NMR studies, and is also applicable for constructs containing the SH2 and SH3 domains of the kinases.

Keywords: Src; Abl; imatinib; tyrosine kinases; biophysical methods; bacterial expression; NMR

Protein tyrosine kinases play a central role in cellular signaling. The dysregulation of the regulatory mechanisms that control protein tyrosine kinase activity is associated with many diseases, particularly cancer. c-Abl and c-Src are closely related non-receptor tyrosine kinases that contain SH2 and SH3 domains in addition to the catalytic tyrosine kinase domain (Thomas and Brugger 1997; Perdognat 2002). Due to the success of the Abl inhibitor imatinib in the treatment of chronic myelogenous leukemia, protein kinase inhibitors have now been established as excellent therapeutic agents in the clinic (Noble et al. 2004; Krause and Van Etten 2005). The high degree of conservation in the sequences of protein kinases and the fact that most of all kinase inhibitors are comprised of ATP, which is the common substrate of protein kinases, makes it difficult to achieve specificity for individual kinases (Capdeville et al. 2002). Considerable effort is therefore being invested in understanding the nuanced differences in conformation and dynamics that distinguish one kinase from the other. Moreover, a substantial fraction of patients undergoing treatment with imatinib develop resistance mutations in the SH3 domain, which render the kinase resistant to imatinib, and understanding the implications of these mutations is of critical importance (Hall et al. 2002; Druker et al. 2003).

Abstract

The Abl and Src tyrosine kinases are key signaling proteins that are of considerable interest as drug targets in cancer and many other diseases. The regulatory mechanisms that control the activity of these proteins are complex, and involve large-scale conformational changes in response to phosphorylation and other regulatory signals. The success of the Abl inhibitor imatinib in the treatment of chronic myelogenous leukemia has shown the potential of kinase inhibitors, but the rise of drug resistance in patients has also shown that drugs with alternative modes of blocking the kinase are needed. The detailed understanding of the mechanisms of protein–drug interaction and drug resistance through biophysical methods demands a method for the production of active protein on the milligram scale. We have developed a bacterial expression system for the kinase domain of c-Abl and c-Src, which allows for the quick expression and purification of active wild-type and mutant kinase domains by coexpression with the YopH tyrosine phosphatase. This method makes practical the use of isotopic labeling of c-Abl and c-Src for NMR studies, and is also applicable for constructs containing the SH2 and SH3 domains of the kinases.

Keywords: Src, Abl; mutably expressed tyrosine kinases; bacterial expression; NMR
Which kinases can we easily express in bacteria?
WHICH KINASES CAN WE EASILY EXPRESS IN BACTERIA?
WHICH KINASES CAN WE EASILY EXPRESS IN BACTERIA?
WHICH KINASES CAN WE EASILY EXPRESS IN BACTERIA?

Which kinases can we easily express in bacteria? (Diagram showing expression levels with colors: Green = Good Expression, Yellow = Fair Expression, Orange = Possibly Optimizable Expression, White = No Expression.)

Available at http://choderalab.org/kinome-expression
Which kinases can we easily express in bacteria?

Table of kinase expression constructs

<table>
<thead>
<tr>
<th>TargetID</th>
<th>Family</th>
<th>Top pdb ID</th>
<th>Yeilding (culture)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPMA_HUMAN</td>
<td>O</td>
<td>OKJ3</td>
<td>OKJ3</td>
</tr>
<tr>
<td>VPMA_HUMAN</td>
<td>O</td>
<td>OKJ3</td>
<td>OKJ3</td>
</tr>
<tr>
<td>VPMA_HUMAN</td>
<td>O</td>
<td>OKJ3</td>
<td>OKJ3</td>
</tr>
<tr>
<td>VPMA_HUMAN</td>
<td>O</td>
<td>OKJ3</td>
<td>OKJ3</td>
</tr>
<tr>
<td>GAK_HUMAN</td>
<td>O</td>
<td>49A9</td>
<td>49A9</td>
</tr>
<tr>
<td>GAK_HUMAN</td>
<td>O</td>
<td>49A9</td>
<td>49A9</td>
</tr>
<tr>
<td>GAK_HUMAN</td>
<td>O</td>
<td>49A9</td>
<td>49A9</td>
</tr>
<tr>
<td>GAK_HUMAN</td>
<td>O</td>
<td>49A9</td>
<td>49A9</td>
</tr>
</tbody>
</table>

An open library of human kinase domain constructs for automated bacterial expression

Daniel L. Parton, Sonya M. Hanson, I. Launoisse-Rodrigue-Laumen, Steven K. Albenes, Scott M. Baken, Chris Jeans, Markus Seeliger, Nicholas Levinson, and John D. Chodera

Department of Pharmacological Sciences, Stony Brook University Medical School, Stony Brook, NY 11794

5 Great Neck Shale Kittinger Graduate School, Memorial Sloan Kittinger Cancer Center, New York, NY 10065

Department of Pharmacy, University of Minnesota, Minneapolis, MN 55455

Available at http://choderalab.org/kinome-expression