REDESIGNING DRUG DESIGN
(AND WHY THIS REQUIRES MULTISCALE MODELING)

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

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

DISCLOSURES:
- Scientific Advisory Board, Schrödinger
BIOLOGICAL MACROMOLECULES ARE THE MOLECULAR MACHINES OF LIFE

enzymes

prolyl isomerase

kinase

nucleic acids

ATP synthase

H⁺

ADP

ATP

unfolded protein

sensors

rhodopsin

turbines

ATP synthase

ADP

ATP

H⁺

signaling molecules

insulin

glucagon

factories

ribosome

structural components

actin

motors

myosin

Illustrations by David Goodsell, courtesy of the RCSB
PERTURBATIONS CAN CAUSE DISEASE
BY DISRUPTING MOLECULAR MACHINES

- Defective binding, catalysis, or regulation
- Broken sensors, channels, and pumps
- Reduced efficacy or production of signaling molecules
- Deficient structural properties
- Defective transport
- Misfolding and aggregation

Illustrations by David Goodsell, courtesy of the RCSB

Jones et al., JMB 2003
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

---

![Graph showing the trend in R&D efficiency](https://example.com/graph.png)

- **Overall trend in R&D efficiency (inflation-adjusted)**

  - 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

**Eroom's Law**
- Overall trend in R&D efficiency (inflation-adjusted)
- FDA tightens regulation post-thalidomide
- First wave of biotechnology-derived therapies
- FDA clears backlog following PDUFA regulations plus small bolus of HIV drugs

**Moore's Law**
- Transistor density over time

---

© 2012 Macmillan Publishers Limited. All rights reserved.
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%

- Cost per launch (out of pocket):
  - $24
  - $54
  - $178
  - $185
  - $235
  - $44
  - $873

- This stage has a 92% failure rate

- ~2-4% overall success rate

- 4.5 years and $219M

WE REGULARLY **DESIGN** PLANES, BRIDGES, AND BUILDINGS ON COMPUTERS

10³ - 10⁶ parts

**WHY NOT** SMALL MOLECULE DRUGS?

< 10² 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?

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.

We use physical modeling and statistical mechanics to build predictive models.

Physical binding constant: $K_d$

Catalytic life cycle: $K_{i,app}$

Cellular pathways: $EC_{50}$
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?

$P + L \leftrightarrow K_d \leftrightarrow PL$
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?

\[ P + L \overset{K_d}{\leftrightarrow} PL \]

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**

**Terms of Interest:**
- entropy
- enthalpy
- conformational heterogeneity
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

if insufficiently accurate, systematically add detail

if accurate enough, systematically remove detail

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

molecular mechanics potential energy
HOW CAN WE **COMPUTE** A BINDING AFFINITY INCLUDING RELEVANT STATISTICAL MECHANICS?

\[ P + L \underset{K_d}{\overset{}{\rightleftharpoons}} PL \]

**dissociation constant** \[ K_d \sim \frac{\tau_{\text{unbound}}}{\tau_{\text{bound}}} \]
How can we compute a binding affinity including relevant statistical mechanics?

ANTON

$50M$ special-purpose supercomputer from D.E. Shaw Research

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**

\[ \Delta G_{\text{bind}} = \Delta G_{P + L} - \Delta G_{P + \emptyset} \]

Multiple simulations of alchemical intermediates

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

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

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

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

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.
- 1.33±0.05 kcal/mol (N=17)  

T4 lysozyme L99A:
- 1.89±0.04 kcal/mol (N=13)  
- 0.6±0.2 kcal/mol (N=3)  

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

JNK3 kinase:
- Anecdotal literature reports of success (publication bias?)
- Calculations are notoriously unreliable. (e.g. SAMPL predictive challenges)

(retrospective RMS error [sample size]
prospective RMS error [sample size]
(not to scale)
Alchemical free energy methods can work reliably in simple systems, but complex systems remain challenging.

**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.

**Pharmacologically relevant**
- JNK3 kinase: large protein, multiple conformations, large drug-like ligands, rotatable bonds, multiple protonation states? tautomers? phosphorylation and activation? peptide substrate? MgCl₂ salt effects?

(not to scale)
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
Markov chain Monte Carlo (MCMC) provides a flexible framework for enhancements.

GHMC, protonation state MC, tautomeric state MC, sidechain NCMC, ... other MC

Can be combined with replica exchange schemes to decrease correlation times.

Can be combined with replica exchange schemes to decrease correlation times.
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?

We can exploit new GPU technologies to reach practical computation times.

$50K 5 \text{ TFLOP/S}

Doubling every 18 months

many CPU-weeks/calculation

$500 5 \text{ TFLOP/S}

Beating Moore’s Law?

overnight on a workstation?
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
Wang, Chodera, Yang, and Shirts. JCAMD 27:989, 2013
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
FAIL FAST, FAIL CHEAP

computational predictions

experimental confirmation
1. Predict how modifications to ligand will change affinity
2. Test experimentally

\[ \Delta G_1 \]
\[ \Delta G_2 \]
\[ \Delta G_3 \]
\[ \Delta G_4 \]
\[ \Delta G_5 \]
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

$\Delta G_1$

$\Delta G_2 +$

$\Delta G_3$
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
AUTOMATE. EVERYTHING.

Automated platform for bacterial cloning, mutagenesis, expression, purification, and binding affinity measurement with 24/7 operational capacity.
ASSAY AUTOMATION CAN CONTROL ERROR
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

TrpSSIN
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?

Easy
Hard
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)

   ![Chemical structures](image)

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 INHIBITORS TESTED HERE THAT ARE CURRENTLY APPROVED FOR USE IN HUMANS.

Supplementary Online

From Lapatinib to ForSUINIB 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 Dn. It became apparent that most of the compounds bound only a relatively small number of kinases with high affinity. Among the exceptions were SUINIB and Dasatinib. Two of the marketed drugs which both bound to kinases tested with Dn and respectively, whereas values for Gefitinib (RESSA) were higher. The figure illustrates the smallmolecule kinases interaction maps for kinase inhibitors. Kinases found to bind are marked with red circles. Where larger circles indicate higher affinity binding interactions are shown. Complete results can be found in the Supplementary Online dataset. The HS dataset is also available through an interactive website: http://www.ambitbio.com. The kinase dendrogram was adapted and is reproduced with permission from Science (http://www.sciencemag.org).
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

Abl: imatinib $\Delta G = -10.9$ kcal/mol
Src: imatinib $\Delta G = -6.2$ kcal/mol
$\Delta \Delta G = 4.7$ kcal/mol
DIFFERENCES IN STABILITIES OF INACTIVE STATES MAY BE RESPONSIBLE FOR ORIGIN OF SOME KINASE INHIBITOR SELECTIVITY

QUANTIFYING CONFORMATION ENERGETICS MAY BE CRUCIAL TO SUCCESSFUL DESIGN OF SELECTIVE KINASE INHIBITORS

ΔG of confinement to binding-competent state

ΔG of binding to binding-competent state

high free energy states (inactive?)

low free energy state (active?)

net ΔG of binding

CAN WE BUILD AN ATLAS OF KINASE STRUCTURES AND ENERGETICS?

Can we build an *Atlas of Kinase Structures and Energetics*?

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

How can we model the statistical dynamics of complex systems?

Brownian diffusion in a model one-dimensional potential
If a separation of timescales exists, we can accurately describe kinetics with a discrete-state Markov model.

Coarse-graining of space into discrete states introduces memory:

\[
\dot{C}(t) = -\int_0^t dt' M(t - t') C(t')
\]

If a separation of timescales exists, first-order rates emerge:

\[
\dot{C}(t) \approx \left[-\int_0^\infty dt' M(t - t')\right] C(t) \equiv KC(t)
\]

This produces a Markov model in continuous or discrete time:

\[
p(t) = e^{Kt} p(0) \quad \text{continuous time (rate matrix)}
\]

\[
p(n\tau) = [T(\tau)]^n p(0) \quad \text{discrete time (transition matrix)}
\]

Because of coarse-graining in space, there must be coarse-graining of time as well!
Once we construct a model for transitions among metastable states, properties of interest can be computed from the model.

The plan:

- **Run simulations**
- **State decomposition**
- **Estimation of rates**
  \[
  \begin{bmatrix}
  k_{11} & k_{12} & \cdots & k_{1N} \\
  k_{21} & \ddots & & \vdots \\
  \vdots & & \ddots & \vdots \\
  k_{N1} & \cdots & \cdots & k_{NN}
  \end{bmatrix}
  \]
- **Property computation**
  - Pathways / kinetics
  - Spectroscopy
  - Populations
  - Errors

Computing properties:

- **Populations**
  \[ p(t) = e^{Kt} p(0) \]
- **Spectroscopy**
  \[ \langle A(t) \rangle_{\rho_0} = \langle A, p(t) \rangle = \sum_{k=1}^{K} e^{-\lambda_k t} \langle A, u_k \rangle \langle u_k, \rho_0 \rangle \]
- **Pathways / kinetics**
  - Transition path theory (TPT)

Additional references:
- Berezhkovskii, Hummer, Szabo. JCP 2009.
Can a Markov model describe simple peptide dynamics?

ACE - ALA - NME
Can a Markov model describe simple peptide dynamics?
Alanine dipeptide has six clearly resolved metastable states
Statistical evolution can still be complex
Statistical evolution can still be complex

1000 NVE trajectories 100 ps in length initiated from canonical distribution within each state

Rate estimate stabilizes after a transient time

If transitions can be described by a phenomenological rate constant, then for $t > \tau_{eq}$

$$T_t = e^{Kt}$$

Compute eigenvalues of transition matrix and relate them to rate matrix

$$T_t h_k = \mu_k h_k = e^{\lambda_k t} h_k$$

Convert eigenvalues to implied timescales

$$t_k = -\lambda_k^{-1} = -\tau / \log \mu_k$$

Timescales must be constant for $t > \tau_{eq}$
A model constructed from short trajectories can reproduce dynamics over long times.
The statistical dynamics can be hierarchical.

Is there some general principle that allows us to deduce this statistical behavior for complex systems?
We don’t need Markovity in coarse-grained dynamics. MSMs approximate slow eigenspace of transfer operator.

**General time-evolution:**

$$u_{t+\tau}(y) = T(\tau) \circ u_t(y) = \frac{1}{\mu(y)} \int_{x \in \Omega} d\mu(x, y; \tau) \mu(x) u_t(x).$$

**Separate evolution operator into slow and fast components:**

$$u_{t+k\tau}(x) = T_{\text{slow}}(k\tau) \circ u_t(x) + T_{\text{fast}}(k\tau) \circ u_t(x)$$

$$= \sum_{i=1}^{m} \lambda_i^k \langle u_i, \phi_i \rangle \psi_i(x) + T_{\text{fast}}(k\tau) \circ u_t(x)$$

**Approximation error can be bounded;**


**Statistical error can be assessed.**

Approximation quality improves with finer partitioning

approximation error bound: 

\[ E(k) \leq \min \{ 2, [m_{\text{skin}} \delta + \eta(\tau)] [a(\delta) + b(\tau)] \} \lambda_2^k \]

\[ a(\delta) = \sqrt{m(k)} \delta \]

\[ b(\tau) = \frac{\eta(\tau)}{1 - \eta(\tau)} (1 - \eta(\tau)^{k-1}) \]

\[ \eta(\tau) = \frac{\lambda_{m+1}(\tau)}{\lambda_2(\tau)} \]


We can approximate the eigenfunctions and discover metastable sets using an adaptive kinetic clustering algorithm.

Maximize metastability while controlling statistical error

\[ M(T) = \frac{1}{M} \sum_{i=1}^{K} T_{ii} \] metastability

Clustering alternates between:
* geometric similarity (heavy-atom RMSD)
* kinetic similarity (microstate transition probabilities)

Based on tremendous quantity of mathematical work by

Ch. Schütte, Deuflhard, Huisinga, Weber, Röblitz (née Kube), Meerbach, Fischer, Junge, Dellnitz, Chodera, Pande, Bacallado, Hummer, Roux…

NMR model of trpzip2

Metastable states at high temperature

trpzip2
[PDBe:1LE1]

Dynamics is dominated by residence within metastable states, with brief excursions to other states.
Structural data on human kinases is incomplete.
ENSEMBLER: AUTOMATING SIMULATIONS AT THE SUPERFAMILY SCALE

Daniel Parton
Postdoc

Patrick Grinaway
PBSB student

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

Sonya Hanson
Postdoc

Kyle Beauchamp
Postdoc

http://biorxiv.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

Vija S. Pande
Stanford University

<table>
<thead>
<tr>
<th>OS Type</th>
<th>Native TFLOPS</th>
<th>x86 TFLOPS</th>
<th>Active CPUs</th>
<th>Active Cores</th>
<th>Total CPUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows</td>
<td>347</td>
<td>347</td>
<td>79790</td>
<td>199354</td>
<td>5634616</td>
</tr>
<tr>
<td>Mac OS X</td>
<td>21</td>
<td>21</td>
<td>7625</td>
<td>55470</td>
<td>184189</td>
</tr>
<tr>
<td>Linux</td>
<td>22</td>
<td>22</td>
<td>7011</td>
<td>32192</td>
<td>795767</td>
</tr>
<tr>
<td>ATI GPU</td>
<td>1019</td>
<td>2150</td>
<td>7174</td>
<td>7174</td>
<td>399830</td>
</tr>
<tr>
<td>NVIDIA GPU</td>
<td>1275</td>
<td>2690</td>
<td>6745</td>
<td>6745</td>
<td>342790</td>
</tr>
<tr>
<td>NVIDIA Fermi GPU</td>
<td>12575</td>
<td>26533</td>
<td>37094</td>
<td>135487</td>
<td>535673</td>
</tr>
<tr>
<td>Total</td>
<td>15259</td>
<td>31763</td>
<td>145439</td>
<td>436422</td>
<td>7892865</td>
</tr>
</tbody>
</table>

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

$\times$ 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

$\sim$ 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.
EVEN LARGE TRAJECTORY DATASETS ARE NOT LARGE ENOUGH

Abl: Presence/absence of discrete state overlap among clones
DISTRIBUTED NAÏVE SIMULATIONS ARE JUST TOO INEFFICIENT

WE NEED ADAPTIVE SAMPLING STRATEGIES
ASSUME THAT WE ARE INTERESTED IN A SPECIFIC RARE EVENT (TRANSITION) A TO B

Naive approach is to run a series of simulations starting in $A$ and just count.

$$\mathbb{P}(A \to B) \leftarrow \frac{\# \{A \to B\}}{\# \{A \to A \cup B\}}$$

This of course will fail for metastable systems since chances of reaching $B$ are too small.
SAMPLING PATHS

The concept of running simulations of specific types can be seen as

*drawing samples from the (sub-)set of trajectories*

We call a (sub-)set of trajectories an ensemble.

Instead of just drawing samples we can employ all the usual tricks to generate samples

*importance sampling*

*multi-ensemble sampling*

*MCMC*
**SCHEME**

1. Define path ensembles
2. Generate samples (trajectories) in each ensemble
3. Reweight
4. Compute target property using joint ensemble

**BENEFITS**

1. No bias in the dynamics
2. Simple reweighting factors
3. Mostly short trajectories
4. No model

**PROBLEMS**

1. Need to know states
2. Need to set interfaces manually
3. Requires good idea of a reaction coordinate
SAMPLING USING MCMC (THE SHOOTING MOVE)

We need samples in specific ensembles or concrete: paths of a specific type

Difficulty is to generate paths with the correct relative probabilities:
   Likely paths should be generated with higher probability

Given an initial path we can use MCMC.
   (1) propose a new trajectory
   (2) accept / reject

This can be done using the *shooting move*.

1. Pick a (random) point in an existing path sampled from a given ensemble.

2. Extend from that point on in one or both directions to generate a proposal trajectory until you reach \( A \) or \( B \)

3. If the new trajectory is not in the ensemble (type of trajectory) reject otherwise use a Metropolis criterion to accept / reject

A Python Framework to run Path Sampling Algorithms

www.openpathsampling.org

✓ Easy to use
  beginners can quickly learn to use it

✓ Easy to extend
  advanced users can use it to develop new methods

✓ Use modern simulation tools
  OpenMM can use GPU acceleration
since ensembles are sets we can use set operations with the fundamental ensembles (and, or, not, ... )
Transition Interface Sampling (TIS)

First frame in A, last frame in A or B, arbitrary length, all other frames not in A or B, at least one frame crosses interface X

Sequential[
    Length(1) & AllIn(A),
    AllOut(A | B) & PartOut(interface),
    Length(1) & AllIn(A | B)
]
ADAPTIVE MARKOV STATE MODEL CONSTRUCTION VIA PATH SAMPLING

Combine advantages of MSMs and path sampling:
1. Use MSM to **discover states** and identify interface sets
2. Use OPS to sample rare events and improve transition matrices

*Could lead to a fully Bayesian MSM approach*
www.openpathsampling.org
A Framework to run Path Sampling Algorithms
Version 1.0 coming soon
Code and data available at http://www.choderalab.org
COLLABORATORS

Stanford
Vijay Pande
OpenMM team
Omnia team
Folding@home team

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

Simon Fraser
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 Edinburgh
Antonia S. J. S. Mey

UC Berkeley
Susan Marqusee
Carlos Bustamante
Christian Kaiser

University of Chicago
Suri Vaikuntanathan

LBNL
Gavin Crooks

Vanderbilt
Joel Tellinghuisen

Hessian Informatics
Kim Branson

Code and data available at http://www.choderalab.org

http://folding.stanford.edu
OMNIA:
OPEN SOURCE, HIGH PERFORMANCE, HIGH USABILITY TOOLKITS FOR PREDICTIVE BIOMOLECULAR SIMULATION.

http://omnia.md