Learning molecular model from simulation and experimental data

Cecilia Clementi

Center for Theoretical Biological Physics
Department of Chemistry & Department of Physics
Rice University, Houston, TX USA

Einstein Visiting Professor
Multiscale Modeling of Biophysical Systems
Department of Mathematics and Computer Science
Freie Universität, Berlin, Germany
A main challenge in biophysics
broad range of interconnected length and time scales

<table>
<thead>
<tr>
<th>System</th>
<th>Length/Time Scale</th>
<th>Number of Atoms</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism</td>
<td>1-10 μm</td>
<td>~10^{20} atoms</td>
<td>Macroscale</td>
</tr>
<tr>
<td>Cell</td>
<td>10-100 nm</td>
<td>~10^{10} atoms</td>
<td>Mesoscale</td>
</tr>
<tr>
<td>System</td>
<td>10-100 nm</td>
<td>~10^{4}-10^{5} atoms</td>
<td>Multiscale</td>
</tr>
<tr>
<td>Biomolecule (Macromolecule)</td>
<td>~10-100 nm</td>
<td>~10^{3}-10^{4} atoms</td>
<td>Multiscale</td>
</tr>
<tr>
<td>Atom</td>
<td>1-10 Å</td>
<td>~1 atom</td>
<td>Quantum Physics</td>
</tr>
</tbody>
</table>

**Quantum Physics**

~1 atom
~1 Å

**Quantum Chemistry**

~10^{-1} - ~10 atoms
~1-10 Å

**Mesoscale Multiscale**

~10^{-3} - ~10^{4} atoms
~1-100 nm

**Reaction Diffusion Simulation**

**All-atom Molecular Dynamics**
Coarse-Graining

**$\mathbb{R}^3_N$**
- coarse-graining in conformation space

**$\mathbb{R}^3$**
- coarse-graining in structural space
Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods

Frank Noé¹ and Cecilia Clementi²
Coarse-Graining

course-graining in structural space
Coarse-graining: what properties should be preserved?

Real experimental observations - Often thermodynamic/macroscopic

“Top-down”

Coarse-grained Model

“Knowledge-based”

Classical, empirical Atomistic Model

“Bottom-up”

Real material - “Fundamental” description

Experimentally Determined Structures

Decreasing resolution/Increasing scale

Coarse-graining in structural space

\[ H = \sum_{\text{bonds}} K_r (r - r_0)^2 + \]
\[ \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2 + \]
\[ \sum_{\text{dihedrals}} K_d (1 + \cos(n(\phi - \phi_0))) + \]
\[ \sum_{i,j} \epsilon_{ij} \left[ 5 \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 6 \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{10} \right] \]

SNARE complex: Results

Results: comparison with experiments

STATE 4:
Unzipped until before ionic layer

STATE 3:
Unzipped up to layer +5

STATE 5:
Unzipped until after ionic layer

• Definition of coarse variables
• Definition of effective energy function (and dynamic equations)
• Incorporation of experimental data
Coarse-grained mapping

\[
\begin{align*}
\text{atomistic representation} & \quad \mathbf{r} \in \mathbb{R}^{3N} \\
\text{CG representation} & \quad \mathbf{x} = \xi(\mathbf{r}) \in \mathbb{R}^{3n} \\
\text{atomistic potential} & \quad V(\mathbf{r}) \\
\text{CG potential} & \quad U(\mathbf{x})
\end{align*}
\]
Coarse-graining with thermodynamic consistency

atomistic representation \( \mathbf{r} \in \mathbb{R}^{3N} \)

CG representation \( \mathbf{x} = \xi(\mathbf{r}) \in \mathbb{R}^{3n} \)

atomistic potential \( V(\mathbf{r}) \)

CG potential \( U(\mathbf{x}) \)

Thermodynamic consistency

\[
U(\mathbf{x}) = -k_B T \ln p^{CG}(\mathbf{x}) + \text{const}
\]

\[
p^{CG}(\mathbf{x}) = \frac{\int \exp \left( -\frac{V(\mathbf{r})}{k_B T} \right) \delta \left( \mathbf{x} - \xi(\mathbf{r}) \right) d\mathbf{r}}{\int \exp \left( -\frac{V(\mathbf{r})}{k_B T} \right) d\mathbf{r}}
\]

GOAL:
Optimize the parameters of a CG model \( U(\mathbf{x}; \theta) \) to satisfy this consistency as best as possible

Force matching

Enforcing the thermodynamic consistency is equivalent to minimize the force matching error:

\[ \chi^2(\theta) = \left\langle \| \xi(F(r)) + \nabla U(\xi(r); \theta) \|^2 \right\rangle_r \]

The force matching error is always > 0 and can be decomposed in two parts:

\[ \chi^2(\theta) = \text{PMF error}(\theta) + \text{Noise} \]

\[ \text{PMF error}(\theta) = \left\langle \| f(\xi(r)) + \nabla U(\xi(r); \theta) \|^2 \right\rangle_r \]

\[ \text{Noise} = \left\langle \| \xi(F(r)) - f(\xi(r)) \|^2 \right\rangle_r \]

\[ f(x) = \left\langle \xi(F(r)) \right\rangle_{r|x} \]

The gradient of the CG potential should be as close as possible to the “mean force”

The PMF error is what should be as close to zero as possible

The noise term is determined solely by the coarse-graining mapping
Coarse-graining as a machine learning problem

Enforcing the thermodynamic consistency is equivalent to minimize the force matching error:

$$\chi^2(\theta) = \left\langle \| \xi(F(r)) + \nabla U(\xi(r); \theta) \|^2 \right\rangle_r$$

Define a loss function to minimize over a Boltzmann-distributed sample:

$$L(\theta; R) = \frac{1}{M} \sum_{i=1}^{M} \| \xi(F(r_i)) + \nabla U(\xi(r_i); \theta) \|^2$$

$$= \| \xi(F(R)) + \nabla U(\xi(R); \theta) \|^2_F$$
Coarse-graining as a machine learning problem

\[ \mathbb{E} [L(\theta; R)] = \text{Bias}^2 + \text{Var} + \text{Noise} \]

\[ \text{Bias}^2 = \| f(X) - \bar{f}(X) \|^2_F \]

\[ \text{Var} = \mathbb{E} \left[ \| \bar{f}(X) + \nabla U(X) \|^2_F \right] \]

\[ \mathbb{E} [L(\theta; R)] = \text{Bias}^2 + \text{Var} + \text{Noise} \]
\[ L(\theta; R) = \frac{1}{M} \sum_{i=1}^{M} \left\| \xi(F(r_i)) + \nabla U(\xi(r_i); \theta) \right\|^2 \]

\[ = \left\| \xi(F(R)) + \nabla U(\xi(R); \theta) \right\|_F^2 \]

\[
\frac{U(x)}{k_B T} = -\ln \left[ \int_{-\infty}^{+\infty} \exp \left( -\frac{V(x, y)}{k_B T} \right) dy \right]
\]
CGnet for Alanine Dipeptide
CGnets

Unregularized CGnet

Regularized CGnet

“Spline” model

\[
p(r_i) \propto \exp\left(-\frac{k_{b,i}(r_i - r_{i0})^2}{2k_BT}\right)
\]

\[
p(\theta_j) \propto \exp\left(-\frac{k_{a,j}(\theta_j - \theta_{j0})^2}{2k_BT}\right)
\]
CGnets for Alanine Dipeptide: results

CGnets for Alanine Dipeptide: results

Coarse-graining of Chignolin folding/unfolding + water
\begin{align*}
p(r_i) & \propto \exp\left(-\frac{k_{b,i}(r_i - r_{i0})^2}{2k_BT}\right) \\
p(\theta_j) & \propto \exp\left(-\frac{k_{a,j}(\theta_j - \theta_{j0})^2}{2k_BT}\right) \\
U_{rep}(r) & = \left(\frac{\sigma}{r}\right)^c
\end{align*}

Regularized CGnet

“Spline” model

Coarse-graining of Chignolin folding/unfolding

Coarse-graining of Chignolin folding/unfolding

• Coarse-graining can be formulated as a machine learning problem

• A neural net naturally captures crucial multi-body effects

• Applications to more complex systems may require additional regularization (more physical constraints)

• Is it possible to design transferable CG models?
What about the dynamics?

Original all-atom Dynamics vs. CGnet Dynamics

First TIC

Second TIC

Third TIC
Consistency

Thermodynamic Consistency:
potential recovers potential of mean force:

\[ F^\xi(z) = - \log \int_{\Sigma_z} \mu(x) J^{-1/2}(x) \, dx \]

Kinetic Consistency:
first few eigenvalues of original dynamics are restored:

\[ \mathcal{L}^\xi \psi_i^\xi = - \kappa_i^\xi \psi_i^\xi \]

\[ \kappa_i \approx \kappa_i^\xi \]

D. Crommelin and E. Vanden-Eijnden, SIAM Multiscale Model. Simul. 9, 1588 (2011)
Overdamped Langevin dynamics with constant diffusion

\[ dx_t = -D \frac{\nabla U}{k_BT} dt + \sqrt{2D} dW_t \]

Generator:

\[ \mathcal{L} = -\frac{D}{k_BT} \nabla U \cdot \nabla + D \Delta. \]

Spectral matching:

\[ \sum_{i=1}^{M} \left\| \mathcal{L} \hat{\psi}_i - \kappa_i \hat{\psi}_i \right\|^2 = 0 \]

\[ \sum_{i=1}^{M} \left\| -\frac{\nabla U}{k_BT} \cdot \nabla \hat{\psi}_i + \Delta \hat{\psi}_i - \frac{\kappa_i}{D} \hat{\psi}_i \right\|^2 = 0 \]
Test first few eigenvalue equations for effective generator in a weak sense:

\[ \theta^* = \arg\min_\theta \sum_{i=1}^{M} \sum_{j=1}^{P} \langle \mathcal{L}_\theta \tilde{\psi}_i^\xi + \tilde{\kappa}_i^\xi \psi_i^\xi, f_j \rangle^2 \]

Requires:
- spectral data \( \tilde{\kappa}_i^\xi, \tilde{\psi}_i^\xi \) (TICA, MSM, …)
- test functions \( f_j \) (user selection)
- parametric model

Output: optimal parameters.
The addition of 100 small random Gaussians to the smooth potential in (a) creates the noisy three-well potential in (b).

The potential learned by using the slow eigenfunctions of the latter is shown in (c):
It successfully encodes the long timescale features as it recovers the position and depth of the main energy minima while smoothing out the local and faster motions.

Example: adjusting diffusion

Assume that thermodynamically consistent potential is available (e.g. from force matching).

Parametric model for the diffusion leads to symmetric generator:

\[ \langle L_\theta^\xi \tilde{\psi}_i, f_j \rangle_\nu = - \int A_\theta^\xi \cdot \nabla_z \tilde{\psi}_i \cdot \nabla_z f_j \, d\nu \]

Using a linear model:

\[ A_\theta^\xi = \left[ \sum_{n=1}^{N} w_n g_n \right] \text{Id} \]

We arrive at a linear regression:

\[ E_1(w) = \| Xw - y \|^2, \]

\[ X_{i,j;n} = -\langle g_n \nabla_z \tilde{\psi}_i \cdot \nabla_z f_j \rangle_\nu, \]

\[ y_{i,j} = -\kappa_i \langle \tilde{\psi}_i, f_j \rangle_\nu. \]

Subject to positivity constraint:

\[ \sum_{n=1}^{N} w_n g_n (Z_{t_k}) \geq a_{min} \geq 0 \]

Example: adjusting diffusion
• Projection to backbone dihedrals.

• Force Matching with 152 Gaussians.

• FM dynamics are too fast by factor 2-3.

• Spectral matching with 104 Gaussians.

• Diffusion is mostly constant.

• Timescales are restored.

Revisiting coarse-graining: outstanding challenges

- Definition of coarse variables
- Definition of effective energy function (and dynamic equations)
- Incorporation of experimental data
Theoretical framework for optimal combination of simulation and experiment

Simulated ensemble from molecular model, $H(\epsilon)$

$$Q(\epsilon) = \mathbb{P} [(e_1(\epsilon), \ldots, e_n(\epsilon)) | \text{Exp}]$$

Quantify agreement as function of the model parameters, $\epsilon$

Optimal model maximizes $Q(\epsilon)$

Experimental observables $(e_1, \ldots, e_n)$, e.g. FRET
Theoretical framework for optimal combination of simulation and experiment

Experimental measurements \((f_1, ..., f_K)\) & uncertainties \((\Delta f_1, ..., \Delta f_K)\)

Initial guess for parameters \(\epsilon^{(0)}\)

1. Simulate MD with \(H^{(0)}(x)\)
2. Estimate equilibrium distribution \(\pi^{(0)}(x)\)
3. Predict experimental observable \(f_k\):
   \[ e_k[\pi] = \sum_i \frac{n_i}{\pi_i} \sum_{j \in S_i} g_k(x_{ij}) \]
4. Model likelihood:
   \[ Q = \mathbb{P}[(e_1[\pi], ..., e_K[\pi]) \mid \text{Exp}] = \prod_k \mathcal{N}(e_k[\pi]; f_k, \Delta f_k) \]
5. Update parameters \(\epsilon^{(1)} = \epsilon^{(0)} + \delta \epsilon\)
6. Reweight stationary distribution to \(\pi_i^{(1)}\)

Application: CG model of FIP35 from “synthetic FRET"

- Generate synthetic FRET distributions from all-atom (DESRES) equilibrium simulations of FIP35
- Learn "optimal" Coarse-Grained model from FRET data
Application: CG model of FIP35 from "synthetic FRET"

Application: CG model of FIP35 from "synthetic FRET"

Application: CG model of FIP35 from "synthetic FRET"

“True” model

0th order model
slowest timescale (folding)
5000 faster than all-atom

Trained CG model
slowest timescale (folding)
500 faster than all-atom

Cross validation: synthetic "Trp fluorescence"

Analysis of parameter space

Ubiquitin as a model system

- $\text{C}_\alpha$-$\text{C}_\beta$ model
- Only native contacts included (219 parameters)
- $^3J(\text{H}_{\text{C}_\alpha}-\text{H}_N)$ spin-couplings as observables (63 values)
- $\Delta f_i = 0.25 \text{ Hz}$

Optimization

- **Model likelihood:**

\[
Q = \prod_{i=1}^{N} N(e_i(\epsilon), f_i, \Delta f_i)
\]

- **Model loss-function:**

\[
L^{(k+1)} = -\ln Q + \alpha \sum_{j=1}^{M} (\epsilon_{j}^{(k+1)} - \epsilon_{j}^{(k)})^2
\]

\(e_i\) calculated observable value

\(f_i\) experimental observable value

\(\Delta f_i\) standard deviation

\(\{\epsilon\}\) model parameters (interaction strength)

---

Results: Ubiquitin with NMR data

To characterize protein systems on a realistic time scale requires a multifaceted approach.

Cecilia Clementi’s research group
http://clementiresearch.rice.edu

Clementi’s group
Dr. Feliks Nüske
Dr. Jiang Wang
Dr. Giovanni Pinamonti
Dr. Fabio Trovato
Eugen Hruska
Wangfei Yang
Nick Charron
Iryna Zaporozhets

previous:
Alex Kluber
Dr. Justin Chen
Dr. Lorenzo Boninsegna (UCLA)
Dr. Fernando Yrazu (Rice)
Dr. Jordane Preto (U Alberta)
Dr. Mary Rohrdanz (MD Anderson)
Dr. Wenwei Zheng (U Arizona)
Dr. Amarda Shehu (GMU)
Dr. Payel Das (IBM)
Dr. Silvina Matysiak (U. Maryland)
Dr. Brad Lambeth (Shell)

$$ NSF $$
$$ Welch Foundation $$
$$ Einstein Foundation $$