Uncovering Dynamical Equations for Coarse-Graining

Sara Rojas [§] * and Lorenzo Boninsegna[§]

[§] Center for Theoretical Biological Physics, Rice University, Houston TX 77005, USA
* University of Houston, Houston TX 77004, USA

Abstract

The enormous amount of molecular dynamics data available calls for an ever-growing need for extracting the macroscopic features while discarding less relevant information; coarse-graining (CG) is a viable strategy to accomplish that.

CG consists in lumping degrees of freedom into effective “beads”. Such low resolution description requires both the coordinates of the newly formed “beads” and their effective interactions. If a suitable CG model is found, interpreting and simulating the low resolution dynamics is easier than it is in the original system. However, current CG models are usually based on the user’s physico - chemical intuition which does not guarantee that the correct low resolution description is recovered; e.g., improper CG grouping could produce unphysical results. Therefore, data-driven CG models are highly desirable: trajectory data already contains the information in which we are interested, we just need to uncover it.

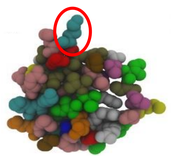
Hereby, we take a first step towards postulating CG equations of motion, as we test a data-based technique which allows to write out the system potential energy function as a sparse linear combination of user-defined functions.
A sparse minimization problem is formulated on Brownian dynamics trajectory data, and solved using cross validation. Preliminary results on a 1d toy model show that our protocol systematically targets an optimally sparse linear combination, which accurately approximates the system energy function. Future perspectives and technique improvement steps are also discussed.

Motivation

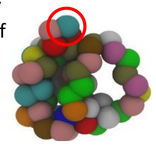
Coarse-graining: procedure used to make a system simpler

Why do we need Coarse-Grained models?

- Highly dimensional systems
- Simulations
- Physical interpretation of system



Simplify the system by reduction of degrees of freedom



High Resolution Description (HRD)

Low Resolution Description (LRD)

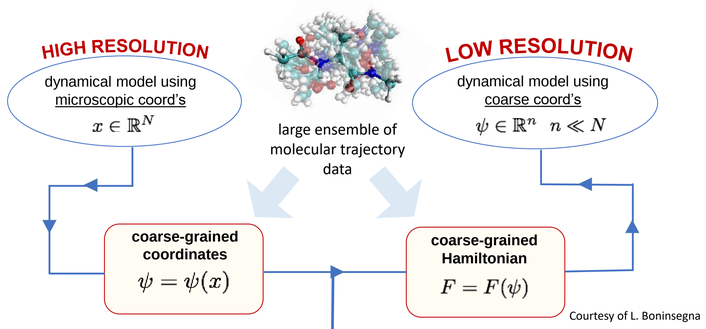
Courtesy of L. Boninsegna

Common Coarse-Graining Approach - Intuition based CG strategy

Complication - Can lead to flawed results

CG Model Requirement – High/Low descriptions need to exhibit the same physical (e.g., long timescale) behavior

Solution - Growing interest in a **data driven**, or **automatic**, CG technique.



Courtesy of L. Boninsegna

HRD Energy Function ≠ LRD Energy Function

The simplified model is described by a **new** energy function, which depends **only** on the remaining degrees of freedom and important HRD interactions.

Propose: Statistical mechanical approach to extract a sparse energy function of the CG model from high resolution trajectory data

Method

A **simple1-D toy model** representing a molecular system was investigated.
The system diffuses at constant temperature, following Brownian Dynamics.
Method: an unknown trajectory was analyzed with the goal of recovering a sparse energy function that generated the dynamics itself.

Brownian Motion (Reduced): $\frac{dx}{dt} = -\frac{dU(x)}{dx} + \eta(t)$ ← Thermal Noise $(\gamma, T, m = 1)$

Ansatz via a library of basis functions: $U(x) = \sum_{q=1}^{K} a_q f_q(x)$ $(t_q = q)$

Loss Function: $\mathcal{L} = \frac{1}{2}\sum_{q=1}^{n}\left\|\frac{dx(q)}{dt} + \frac{dU(x(q))}{dx}\right\|^2$

Minimizing the Loss Function:
- Take partial derivative with respect to each coefficents and set to zero.
- Solve for coefficents; optimal coefficients.

$\frac{\partial \mathcal{L}}{\partial a_k} = 0$

$\mathbb{X}$ Matrix: Time Steps ⇓, Basis Functions ⇒

$$\begin{bmatrix} f_1(x(0)) & \dots & f_K(x(0)) \\ \vdots & & \vdots \\ f_1(x(n)) & \dots & f_K(x(n)) \end{bmatrix}$$

$\mathbb{Y}$ Matrix: Time Steps ⇓

$$\begin{bmatrix} \frac{x(1)-x(0)}{\Delta t} \\ \vdots \\ \frac{x(n)-x(n-1)}{\Delta t} \end{bmatrix}$$

Optimal Coefficients: $\hat{a} = min\|\mathbb{Y} + \mathbb{X}a\|^2 - \lambda|a|$ ← Sparsity Constraint

Stepwise Sparse Regressor (SSR) [4]

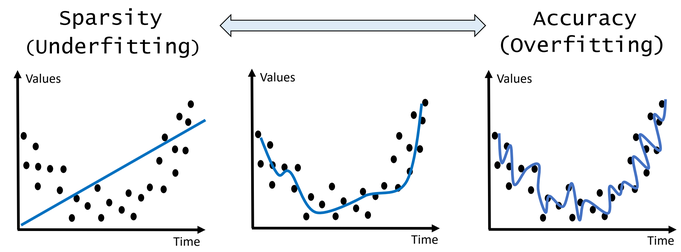
- Solve a standard least square regression
- Smallest coefficient is set to zero
- Solve remaining regression
- Smallest coefficient is set to zero
- Solve remaining regression

“Law of Parsimony”

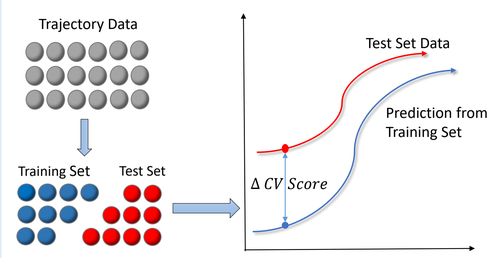
Any value of sparsity can be achieved, however the solutions must be statistically meaningful.

Cross validation is used to ensure values are statistically meaningful, and targets the one solution which optimally balances sparsity and accuracy in data description.

Cross Validation : “Goldilocks principle”



Cross Validation Score

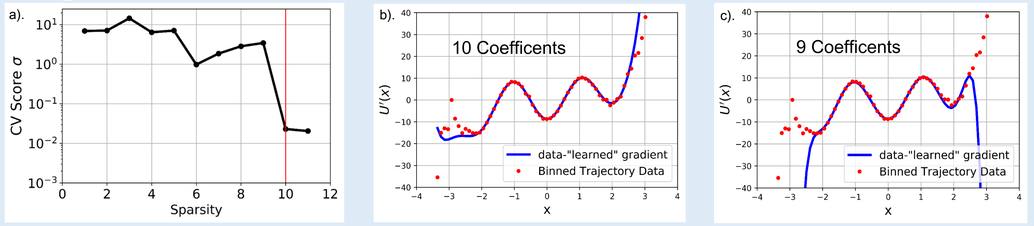


K-fold Cross Validation:

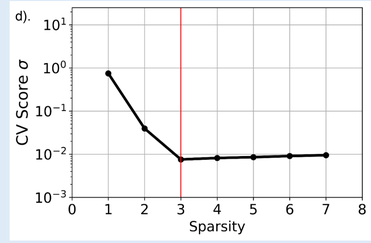
- Splits data into k folds
- Regression is run on each training set and its prediction is compared to test set
- Deviations between test set values and predictions are averaged; cross validation score

Advantages:
- Fully data-based
- Assesses algorithm predictive ability
- Allows algorithm access to all information from data to build predictions

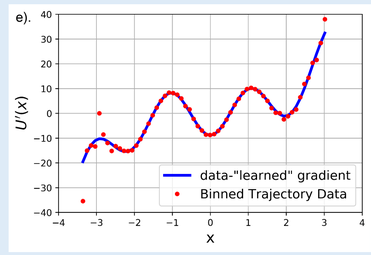
Algorithm Performance

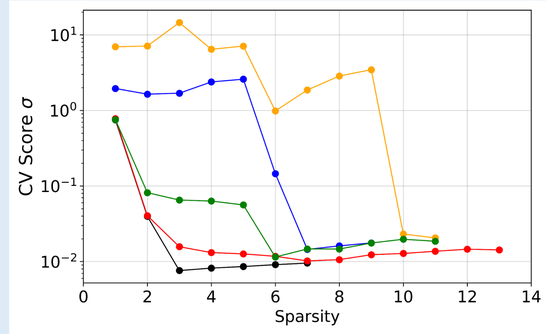


a) CV score as a function of solution sparsity for a given 11 basis function sub-library; the (slightly sparse) 10 term optimal solution is indicated by a red vertical line and is plotted in panel (b) on the top of trajectory data. A 9 coefficient sparser solution is plotted in panel (c) and shows how increasing sparsity influences a solution statistical meaning; specifically, such solution, while sparser, is unphysical.



(d) CV score plot as a function of solution sparsity, for a carefully chosen library of 7 basis functions. A red vertical line identifies the optimal solution which is both accurate and sparse (only 3 terms out of 7 survive): sparser solutions are more parsimonious, but inaccurate (as shown by larger values in CV score).
The optimal solution is plotted on the top of input trajectory data in panel (e), and the agreement is satisfactory. We noticed that the three basis functions “surviving” the SSR thresholding are the same functions ($x^2, x^3, \cos(3x)$) which were incidentally used to generate the trajectory data in the first place.





Cross Validation Score deviation plotted against sparsity, shown for 5 different sub-libraries of basis functions. The optimal solution is strongly dependent on the library composition. However, each accurately reproduces the potential energy (data not shown).

- Algorithm recovered a potential energy form containing:
  - either the analytic functions
  - or an accurate functional form for the energy without the analytic functions
- Accuracy of results highly depends on the sub-libraries provided
- Up to the user to determine if the solution is meaningful
- The **quality** of the sample is extremely important
  - Must be a long simulation
  - Trajectory run with a few stray data points can lead the algorithm in the wrong direction

Outlook

- Modify the algorithm to reconsider basis functions set to zero in previous iterations.
- Algorithm can suppress an analytic function early in the process, and stay in a local minima.
- Algorithm to be applied to more complex systems
  - 2D Toy models
  - Real macromolecular systems (e.g, integrate out water degrees of freedom and factor them in the solute only)

Acknowledgments

Funding Agency:
National Science Foundation:
NSF PHY-1427654

Mentor:
Lorenzo Boninsegna

References

1. Saunders, Marissa G., and Gregory A. Voth. “Coarse-Graining Methods for Computational Biology.” *Annual Review of Biophysics*, vol. 42, no. 1, 2013, pp. 73–93., doi:10.1146/annurev-biophys-083012-130348.
2. Noid, W. G. “Perspective: Coarse-Grained Models for Biomolecular Systems.” *The Journal of Chemical Physics*, vol. 139, no. 9, 2013, p. 090901., doi:10.1063/1.4818908.
3. Clementi, Cecilia. “Coarse-Grained Models of Protein Folding: Toy Models or Predictive Tools?” *Current Opinion in Structural Biology*, vol. 18, no. 1, 2008, pp. 10–15., doi:10.1016/j.sbi.2007.10.005.
4. Boninsegna, Lorenzo, et al. “Sparse Learning of Stochastic Dynamical Equations.” *The Journal of Chemical Physics*, vol. 148, no. 24, 2018, p. 241723., doi:10.1063/1.5018409.
5. Brunton, Steven L., et al. “Discovering Governing Equations from Data by Sparse Identification of Nonlinear Dynamical Systems.” *Proceedings of the National Academy of Sciences U.S.A.*, vol. 113, no. 15, 2016, pp. 3932–3937., doi:10.1073/pnas.1517384113.