Supporting Information for
An Automatic Method for Identifying Reaction Coordinates in Complex Systems

Ao Ma and Aaron R. Dinner

December 1, 2004

1 Genetic neural network method

For a database of physical variables and committor (p_B) values, neural networks are used to determine the functional dependence of p_B on combinations of coordinates, and a genetic algorithm selects the combination that yields the best fit. Each of the components in the search method is discussed, followed by details concerning construction of the database.

1.1 Neural network

Artificial neural networks are widely used for model-free non-linear fitting. An example of the specific form employed in the present study[1–3] is shown in the Fig. 1 of the main text. There are three layers: an input layer, a hidden layer, and an output layer. The values of the nodes in the input layer are those of the physical variables of interest. The values of the nodes in the remaining two layers are those of the sigmoid function

$$f = \frac{1}{1 + \exp(-\theta - \sum_i w_i p_i)} \tag{1}$$

where p_i are the values of the elements in the previous layer, the summation runs over those elements, and the w_i are corresponding weights (represented by lines in Fig. 1). The output is a predicted p_B. Training the neural networks consists of using a scaled conjugate gradients method to vary the w_i to minimize the root mean square (RMS) error in committor values:

$$\text{RMS error} = \sqrt{\frac{1}{M}\sum_i (p_{B,i}^{GNN} - p_{B,i})^2}. \tag{2}$$

Here, the summation runs over the M samples in the database, $p_{B,i}^{GNN}$ is the predicted committor value of configuration i, and p_B is the actual one. As detailed in the main text, in each application, the weights were optimized with a training set and the quality of the fit was evaluated with an independent test set.

1.2 Genetic algorithm

As illustrated in the main text, it is important to be able to evaluate combinations of a large number of physical variables. For example, in the present study, there are ${}_{5812}C_3 = 32,704,036,820$ possible three-descriptor models in the explicit solvent case. Consequently, exhaustive enumeration is not feasible, and we use a genetic algorithm to search the space of coordinate combinations. In this procedure, individuals in a generation of size L consist of combinations of descriptors, and their fitness is determined by the RMS error of the correpsonding trained neural network (Eq. 2). At each cycle of the genetic algorithm, the best l individuals are kept, the remainder are discarded, and $L - l$ new individuals (children) are created by duplicating existing combinations (asexual reproduction by parents) and changing (mutating) one descriptor in the set. In the present study, optimization was performed for 20 to 30 generations of $300 \leq L \leq 2000$, depending on the number of physical variables used as inputs to the neural networks. The genetic algorithm was terminated when the best model persisted for several generations. The results were not very sensitive to L.

1.3 Sampling p_B with uniform distribution

To ensure that the fitting procedure does not reproduce one range of p_B at the expense of others, it is necessary to construct the database in such a way that the distribution of committor values is approximately uniform. To this end, the following procedure was employed. Each trajectory of N saved structures harvested by transition path sampling (see below) was divided into intervals of roughly $\sqrt{N}$ (structures were saved either every 10 or 20 fs, such that $121 \leq N \leq 331$ depending on the path length). Then, intervals were searched sequentially to obtain p_B in each bin of width 0.1.

For example, for $0 \leq p_B < 0.1$, p_B for the $\sqrt{N}$-th structure was evaluated with a small number of trials (10 or 20), and, if that p_B was greater than the lower bound of the target range ($p_B > 0$), structures in that interval were tested one by one with the same number of trials. For the structures with $0 \leq p_B < 0.1$, p_B was re-evaluated with a large number of trials (typically, 100, as detailed in the main text) until one that maintained p_B in the target range was found. Then, the target range was increased to correspond to the next bin ($0.1 \leq p_B < 0.2$ in the example), and the procedure described immediately above was repeated for the same interval. If the initially tested endpoint had p_B less than the lower bound of the target range or no structure with p_B in the target range was found, the interval considered was increased (to structures indexed $\sqrt{N}+1$ to $2\sqrt{N}$ in the example). The search terminated when either p_B values in all the bins were obtained or the end of the trajectory was reached.

1.4 Physical variables to characterize solvent

Here, additional details are provided for the calculation of selected coordinates described in the main text.

1.4.1 Grid-based water densities

Angularly-restricted radial solvent distribution functions were calculated around solute atoms. To this end, three bonded atoms (denoted A, B, and C; see Table 1) were used to construct local right-handed Cartesian coordinate systems. Specifically, the origin was placed at A, the z axis was placed along the line joining

A and B, and the xz plane was defined to contain the three points. Data were binned radially into three spherical shells: $r \leq 3.4$ Å, $3.4 < r \leq 5.8$ Å, and $5.8 < r \leq 8.0$ Å. These regions were divided into four even intervals in $\cos\theta$ and eight even intervals in ϕ, yielding 96 grid cells around each solute atom A. Intervals of alternative sizes did not increase the likelihood of selecting these descriptors in the GNN procedure. For each grid cell, numbers of solvent atoms (oxygen and hydrogen, separately or together) and total charge (with $-0.834e^-$ for oxygen and $0.417e^-$ for hydrogen) were evaluated.

1.4.2 Torque calculations

Torques exerted by the solvent around selected bonds were determined. The bonds considered were the four corresponding to non-trivial dihedral rotations of the peptide backbone: 1C-2N, 2N-2C$_\alpha$, 2C$_\alpha$-2C, and 2C-3N. For each bond between atoms A and B, we computed the solvent-associated force on solute atom C ($\mathbf{F}_s^C$) and then the torque

$$\mathcal{N}_{\text{A-B}} = (\mathbf{F}_s^C \times \mathbf{r}_{BC}) \cdot \hat{\mathbf{r}}_{AB}.$$

In each case, in addition to the total non-bonded force, separate Coulomb electrostatic and van der Waals terms were considered. Also, the forces were computed either with all solvent molecules or only those in spherical shells around C of $r < 3.4$ Å, $r < 4.5$ Å, $r < 6.0$ Å, or $r < 8.0$ Å. Finally, because torques around a given bond are additive, they were grouped and summed as indicated in Table 2.

2 Dynamic simulation details

2.1 Model

We represented the alanine dipeptide with the CHARMM polar hydrogen topology and parameter sets[4,5] and the explicit water molecules with a modified form of TIP3.[5,6] In the gas phase simulations, no cutoffs were introduced. In the explicit solvent simulations, a group-based switching function was used to truncate the interactions by scaling the potential between 6.0 Å and 8.0 Å,[4] consistent with the short-range spherical cutoffs employed in the initial parameterization of the water model.[6] In the implicit solvent simulations, the cutoffs employed were those designated for the ACE2 energy term.[7–10]

2.2 Transition path sampling

The dynamics of the alanine dipeptide isomerizations were simulated with the leap frog Verlet algorithm[11] with a time step of 1 fs. The lengths of bonds to hydrogen atoms were constrained with SHAKE.[12,13] Due to the relative simplicity of the system, we were able to generate initial paths for the reactions studied by guessing approximations to the transition states and firing a large number ($\tilde{1}000$) of random trajectories from those configurations until paths with endpoints in the defined basins were found by chance. Subsequent paths were generated by making shooting moves[14,15] with momentum perturbations of 20% in the vacuum and implicit solvent cases and 4% in the explicit solvent case. Paths were saved every 100 steps of the transition path sampling procedure.

2.3 Umbrella sampling

To harvest putative transition states with $p_B^{GNN} \approx 0.5$, we used the Monte Carlo module[16] in CHARMM.[4] The allowed moves for the peptide were single atom displacements of up to 0.075 Å and torsion rotations of up to 30°. Water molecules were simultaneously translated up to 0.25 Å and rotated around random axes up to 25°. These three types of moves were chosen with relative frequencies of 12:4:125, respectively. One MC step corresponds to a single application of one of the above moves. In the vacuum and implicit solvent simulations, the system was equilibrated for 10^7 MC steps and then sampled every 10^3 MC steps for 4×10^7 MC steps at 300 K. The explicit solvent simulations were more computationally costly; the system was equilibrated for 10^6 MC steps and then sampled every 10^3 MC steps for 10^7 MC steps.

References

1. So, S.-S.; Karplus, M. *J. Med. Chem.* **1996,** *39,* 1521-1530.
2. So, S.-S.; Karplus, M. *J. Med. Chem.* **1996,** *39,* 5246-5256.
3. Dinner, A. R.; So, S.-S.; Karplus, M. *Adv. Chem. Phys.* **2002,** *120,* 1.
4. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comp. Chem.* **1983,** *4,* 187-217.
5. Neria, E.; Fischer, S.; Karplus, M. *J. Chem. Phys.* **1996,** *105,* 1902-1921.
6. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983,** *79,* 926-935.
7. Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996,** *100,* 1578-1599.
8. Schaefer, M.; Bartels, C.; Leclerc, F.; Karplus, M. *J. Comp. Chem.* **2001,** *22,* 1857.
9. Calimet, N.; Schaefer, M.; Simonson, T. *Proteins* **2001,** *45,* 144.
10. Schaefer, M.; Karplus, M. *J. Mol. Bio.* **1998,** *284,* 835-848.
11. Hockney, R. W.; Eastwood, J. W. *Computer Simulation Using Particles;* McGraw-Hill: New York, 1981.
12. van Gunsteren, W. F.; C., B. H. J. *Mol. Phys.* **1977,** *34,* 1311-1327.
13. Ryckaert, J.-P.; G., C.; C., B. H. J. *J. Comp. Phys.* **1977,** *23,* 327-341.
14. Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002,** *53,* 291-318.
15. Dellago, C.; Bolhuis, P. G.; Geissler, P. L. *Adv. Chem. Phys.* **2002,** *123,* 1.
16. Dinner, A. R. *Monte Carlo Simulations of Protein Folding,* Thesis, Harvard University, 1999.

Table 1: Atoms used to define local coordinate systems for angularly-restricted solvation shells

Index	A	B	C
1	2N	$2C_\alpha$	2C
2	2O	2C	$2C_\alpha$
3	3N	2C	$2C_\alpha$
4	$1CH_3$	1C	2N
5	1O	1C	2N
6	2H	2N	$2C_\alpha$
7	$2C_\alpha$	2C	3N
8	$2C_\beta$	$2C_\alpha$	2C
9	2C	$2C_\alpha$	2N
10	1C	2N	$2C_\alpha$
11	3H	3N	2C
12	$3C_\alpha$	3N	2C
13	$2C_\alpha$	2C	2O
14	$2C_\alpha$	$2C_\beta$	2N
15	$2C_\alpha$	2C	$2C_\beta$
16	$2C_\alpha$	2C	2N
17	2C	$2C_\alpha$	2O
18	2C	2N	$2C_\beta$
19	$2C_\alpha$	1C	3N
20	2C	2N	3N
21	2C	2N	3H
22	2C	2N	1O
23	2C	1O	3H
24	3N	1O	3H
25	2C	2O	2N
26	2N	2C	2O
27	3N	2N	$2C_\alpha$
28	3N	2N	2C

Table 2: Groupings employed in torque calculations. Forces between the solvent and the indicated atoms were calculated. The total torque arising from each group of atoms around each specified bond was then calculated.

Groups of Atoms	Bonds
1C, 1O	2N-2C_α, 2C_α-2C, 2C-3N
2N, 2H	2C_α-2C, 2C-3N
2C, 2O	1C-2N, 2N-2C_α
3N, 3H	1C-2N, 2N-2C_α, 2C_α-2C
3N, 3H, 3C_α	1C-2N, 2N-2C_α, 2C_α-2C
1C, 1O, 2N, 2H	2C_α-2C, 2C-3N

Table 3: Linear regression statistics for models identified in the main text.

Reaction	n	Inputs	Input RMS		Model RMS	
			Train	Test	Train	Test
$C_{7eq} \rightarrow C_{7ax}$ (100 trials for p_B)	1	1C-2N-2C$_\alpha$-2C (ϕ)	0.138	0.138	0.138	0.138
	2	1C-2N-2C$_\alpha$-2C (ϕ)	0.138	0.138	0.136	0.137
		1O-1C-2N-2C$_\alpha$ (θ)	0.293	0.295		
	3	1C-2N-2C$_\alpha$-2C (ϕ)	0.138	0.138	0.136	0.138
		1O-1C-2N-2C$_\alpha$ (θ)	0.293	0.295		
		2N-2C$_\alpha$-2C-3N (ψ)	0.213	0.272		
$C_{7eq} \rightarrow C_{7ax}$ (400 trials for p_B)	1	1C-2N-2C$_\alpha$-2C (ϕ)	0.160	0.160	0.160	0.160
	2	1C-2N-2C$_\alpha$-2C (ϕ)	0.160	0.160	0.160	0.132
		1O-1C-2N-2C$_\alpha$ (θ)	0.267	0.205		
	3	1C-2N-2C$_\alpha$-2C (ϕ)	0.160	0.160	0.160	0.132
		1O-1C-2N-2C$_\alpha$ (θ)	0.267	0.205		
		2C$_\beta$-2C-2C$_\alpha$-2N (γ)	0.276	0.209		
$C_{7eq} \rightarrow \alpha_R$ (vacuum)	1	2N-2C$_\alpha$-2C-3N (ψ)	0.125	0.125	0.125	0.125
	2	2N-2C$_\alpha$-2C-3N (ψ)	0.125	0.125	0.114	0.112
		2O-2C-3N-3C$_\alpha$ (θ')	0.290	0.283		
$C_{7eq} \rightarrow \alpha_R$ (explict solvent and instant coordinates)	1	2C$_\beta$-2C$_\alpha$-2C-2O (ψ')	0.177	0.156	0.177	0.156
	2	2C$_\beta$-2C$_\alpha$-2C-2O (ψ')	0.177	0.156	0.174	0.149
		$\mathcal{N}^{3H}_{1C\text{-}2N}$	0.225	0.210		
	3	2C$_\beta$-2C$_\alpha$-2C-2O (ψ')	0.177	0.156	0.172	0.141
		$\mathbf{r}_{2H\text{-}2C_\beta}$	0.256	0.252		
		$\mathcal{N}^{3H}_{1C\text{-}2N}$	0.225	0.210		
$C_{7eq} \rightarrow \alpha_R$ (explict solvent and average coordinates)	1	2C$_\beta$-2C$_\alpha$-2C-2O (ψ')	0.177	0.156	0.177	0.156
	2	$\langle \mathbf{r}_{2H\text{-}3H} \rangle_{\text{solute}}$	0.196	0.182	0.169	0.139
		$\langle \mathcal{N}^{3C_\alpha}_{1C\text{-}2N} \rangle_{\text{solute}}$	0.179	0.160		
	3	$\langle \mathbf{r}_{2H\text{-}3H} \rangle_{\text{solute}}$	0.196	0.182	0.166	0.135
		$\langle \mathcal{N}^{3C_\alpha}_{1C\text{-}2N} \rangle_{\text{solute}}$	0.179	0.160		
		$\langle \mathcal{N}^{3N}_{1C\text{-}2N} \rangle_{\text{solvent}}$	0.201	0.178		
	4	$\langle \mathbf{r}_{2H\text{-}3H} \rangle_{\text{solute}}$	0.196	0.182	0.165	0.132
		$\langle \mathcal{N}^{3C_\alpha}_{1C\text{-}2N} \rangle_{\text{solute}}$	0.176	0.160		
		$\langle \mathcal{N}^{3N}_{1C\text{-}2N} \rangle_{\text{solvent}}$	0.201	0.178		
		$\langle \mathbf{r}_{1O\text{-}2H} \rangle_{\text{solute}}$	0.256	0.251		
$C_{7eq} \rightarrow \alpha_R$ (implicit solvent)	1	2N-2C$_\alpha$-2C-2O (ψ'')	0.191	0.184	0.191	0.184
	2	2N-2C$_\alpha$-2C-3N (ψ)	0.190	0.184	0.130	0.123
		2O-2C-3N-3C$_\alpha$ (θ')	0.287	0.289		
	3	2N-2C$_\alpha$-2C-3N (ψ)	0.191	0.184	0.093	0.100
		2O-2C-3N-3C$_\alpha$ (θ')	0.287	0.289		
		2C$_\beta$-2C-2C$_\alpha$-2N (γ)	0.287	0.289		