Classification of Protein-protein Association Rates Based on Biophysical Informatics

Kalyani Dhusia, and Yinghao Wu

Supporting Information
Model validation for the Monte-Carlo simulations of protein-protein association

In order to evaluate the reliability of our residue-based kinetic Monte-Carlo simulation method, we systematically adjusted different parameters in the simulation. We first tested the effects of different energetic contributions on regulating the protein-protein association. The scoring function used in our Monte-Carlo simulation contains two energy terms. One is the electrostatic interaction and the other is the hydrophobic interaction between two interacting proteins in the system. We have carried out two specific tests. In one test we turned off the contribution of the electrostatic interaction by setting its weight to 0, while in the other test we turned off the contribution of the hydrophobic interaction. We further compared their simulation results with the situation in which both electrostatic and hydrophobic interactions are on. In detail, the protein complexes 2VLN and 2VLR were used as two test systems. The experimental association rate of 2VLN is $1 \times 10^8 \text{M}^{-1}\text{s}^{-1}$. This rate constant is at the upper bound of diffusion-limited region, in which protein-protein associations are dominated by long-range electrostatic interactions. In contrast, the association of 2VLR is much slower, with a rate of $5 \times 10^4 \text{M}^{-1}\text{s}^{-1}$. For both complexes, ten different values of distance cutoff from 16Å to 25Å were test for the system without electrostatic interaction, as well as for the system without hydrophobic interaction. Under each condition, $10^3$ independent simulation trajectories were carried out and the association probability was then calculated. These calculated probabilities are summarized in Figure S2 as a function of distance cutoff. The systems without hydrophobic interaction are shown by red circles. The systems without electrostatic interaction are shown by blue triangles, while the systems with both electrostatic and hydrophobic interactions are shown by black squares. As shown in Figure S2a, the association of 2VLN has been completely diminished when the electrostatic interaction was turned off. On the other hand, there are few impacts on its association when the hydrophobic interaction was turned off. This result indicates that the association of complex 2VLN is dominated by its electrostatic interaction. Different from 2VLN, the elimination of either electrostatic or hydrophobic interaction did not fully block the association of complex 2VLR, as shown in Figure S2b. This result indicates that its association is regulated by the combination of both effects. Our simulations therefore show the evidence that protein-protein association is controlled by both electrostatic and hydrophobic interactions, and in a case-dependent manner. We also suggest that while the domination of long-range electrostatic interaction leads into fast association, short-range hydrophobic interaction plays a more important role in slow association.

In the next step, we tested the calculation of association probability by changing different numbers of simulation trajectories. In our original study, $10^3$ independent simulation trajectories were carried out under different values of distance cutoff. Larger numbers of trajectories were further tried. In detail, the protein complex 2VLN was used as a test model. For each distance cutoff from 16Å to 25Å, three scenarios were generated. In the first scenario, $10^3$ independent simulation trajectories were performed; in the second scenario, $2 \times 10^3$ independent simulation trajectories were performed; and in the third scenario, $5 \times 10^3$ independent simulation trajectories were performed. Under each value of distance cutoff and each scenario, the numbers of trajectories in which encounter complexes were successfully formed have been recorded. The results are summarized in Figure S3a. The black squares in the figure correspond to the scenario with $10^3$ trajectories, while the red circles and blue triangles correspond to the scenarios with $2 \times 10^3$ and $5 \times 10^3$, respectively. These numbers of successful association have further been converted to the association probabilities by dividing with the total numbers of trajectories. The derived association probabilities for all three scenarios are shown in Figure S3b. The figure
indicates that the profiles of association probability from different scenarios are very close to each other. This result suggests that the calculations of association probability were converged. Further increase of simulation trajectories did not affect the outcome. As a result, $10^3$ simulation trajectories are sufficient to obtain statistically meaningful results of association probabilities.

We further calibrated the acceptance rate in our Monte-Carlo simulations. The calibration was carried out by adjusting the simulation temperature which is embedded in the Metropolis Criteria of the algorithm. In detail, the protein complex 2VLN was used as a test model. Five different values of temperature were tried. They are $0.2T_0$, $0.5T_0$, $T_0$, $2T_0$, and $5T_0$, in which $T_0$ is the room temperature used in our original model. For each temperature, distance cutoff was varied from 16Å to 25Å, while $10^3$ independent simulation trajectories were generated under each distance cutoff value. We calculated the acceptance rate of Monte-Carlo movements from the 1000 simulation steps within each of these trajectories. The average values of these calculated acceptance rates were plotted together with their standard deviation in Figure S4a as a function of simulation temperature. Not surprisingly, the figure shows that the acceptance rate becomes higher when the temperature increases, while the standard deviation of their distributions becomes smaller. Under the temperature of $T_0$, the overall acceptance rate equals 0.87. We also plotted the association rates as a function of distance cutoff under different simulation temperatures in Figure S4b. The association profiles in the figure shows that the overall probabilities drop as the temperature increases. Under low temperature, the energetically unfavored moves are more likely to be rejected. As a result, the Monte-Carlo acceptance rate becomes low but the proteins are more likely to find their native-like binding conformation which leads to the increase of association probabilities. In contrast, under high temperature, the energetically unfavored moves are more likely to be accepted. As a result, the Monte-Carlo acceptance rate becomes high but the proteins are less likely to find their native-like binding conformation which leads to the decrease of association probabilities.

Finally, we changed the lengths of each time step in the simulations. In our original model, each trajectory consists of $10^3$ steps and each time step is 0.01 ns, so that the total simulation time duration for each trajectory is 10 ns. In addition to this setting, two scenarios were further tested. In the first scenario, we changed the length of each simulation time step into 0.1ns. In order to keep the total time duration of each trajectory unchanged, the number of simulation steps was reduced from 1000 to 100. In the other scenario, we changed the length of each simulation time step into 0.001ns. In order to keep the total time duration of each trajectory unchanged, the number of simulation steps was increased from 1000 to 10000. All three scenarios, including the original setting, were applied to the test model 2VLN. In detail, within each scenario, $10^3$ independent simulation trajectories were performed for each distance cutoff from 16Å to 25Å and the association probabilities were further calculated. The derived association probabilities for all three scenarios are summarized in Figure S5. The simulation results from the system which time step equals 0.001ns are plotted by red circles. The simulation results from the system which time step equals 0.01ns are plotted by black squares. The simulation results from the system which time step equals 0.1ns are plotted by blue triangles. All three scenarios show that the association probabilities drop as the distance cutoff increases. However, with larger time step (0.1ns), much lower association probabilities were derived, indicating that the simulation with only 100 steps for each trajectory is not sufficient. On the other hand, the association profiles derived from the other two time-steps are very close, indicating that the setting of the time step in our original model is sufficient for conformational sampling.
Figure S1: The comparison of the prediction results from our neural-network-based classification method (black) with the prediction results from random estimation (red) for sensitivity (a), specificity (b), precision (c) and accuracy (d), respectively.
Figure S2: We tested the effects of different energetic contributions on regulating the protein-protein association.
Figure S3: We tested the calculation of association probability by changing different numbers of simulation trajectories.
Figure S4: We calibrated the acceptance rate in our Monte-Carlo simulations by adjusting the simulation temperature.
Figure S5: We tested the calculation of association probability by changing the lengths of each time step in the simulations.
Figure S6: The number of intermolecular contacts between residues at native binding interfaces was used as an additional dimension of inputs to the neural-network classification. The classification outcomes were compared with the original results in (a). We further plotted the correlation between the native contacts and the experimentally measured association rates for all 96 protein complexes in the benchmark (b).
|                  | # of samples | Mean | S. D. | Minimum | Median | Maximum |
|------------------|--------------|------|-------|---------|--------|---------|
| **Neural Network** |              |      |       |         |        |         |
| Sensitivity      | 100          | 0.65 | 0.030 | 0.57    | 0.66   | 0.73    |
| Specificity      | 100          | 0.67 | 0.025 | 0.63    | 0.67   | 0.72    |
| Precision        | 100          | 0.71 | 0.019 | 0.67    | 0.71   | 0.75    |
| Accuracy         | 100          | 0.66 | 0.020 | 0.63    | 0.65   | 0.70    |
| **Random Guess** |              |      |       |         |        |         |
| Sensitivity      | 100          | 0.49 | 0.067 | 0.32    | 0.50   | 0.66    |
| Specificity      | 100          | 0.49 | 0.067 | 0.28    | 0.48   | 0.62    |
| Precision        | 100          | 0.55 | 0.042 | 0.44    | 0.54   | 0.64    |
| Accuracy         | 100          | 0.49 | 0.043 | 0.39    | 0.48   | 0.59    |

**Table S1:** The comparison of statistical distributions between results from neural network and results from random estimation.
| PDB  | Partner  | Partner ko(exp) | Log10(ko) | Category | Pred_P(N) | T/F | P(N) (2) | P(N) (3) | P(S) (2) | P(S) (3) | P(S) (4) | P(S) (5) | P(S) (6) | P(S) (7) | P(S) (8) | P(S) (9) |
|------|---------|----------------|-----------|----------|-----------|-----|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| 1A22 | A       | B              | 1080      | 4.8712   | -1       | 0.368 | 0.031    | 0.215    | 0.049    | 0.069    | 0.026    | 0.019    | 0.004    | 0.007    | 0.002    | 0.002    |
| 1A3C | A       | B              | 1060      | 4.9025   | -1       | 0.368 | 0.031    | 0.215    | 0.049    | 0.069    | 0.026    | 0.019    | 0.004    | 0.007    | 0.002    | 0.002    |
| 1B82 | A       | D              | 4.137     | 6.5347   | -1       | 0.0898 | 0.172    | 0.139    | 0.096    | 0.118    | 0.046    | 0.025    | 0.044    | 0.003    | 0.007    | 0.007    |
| 1B9U | A       | D              | 3.717     | 7.5682   | -1       | 0.1097 | 0.252    | 0.219    | 0.141    | 0.148    | 0.103    | 0.057    | 0.038    | 0.026    | 0.015    | 0.002    |
| 1CH1 | A       | D              | 5.893     | 8.9682   | -1       | 0.3708 | 0.362    | 0.218    | 0.16    | 0.102    | 0.032    | 0.046    | 0.018    | 0.01    | 0.021    | 0.021    |
| 1H4I | A       | B              | 5.0129    | 6.9329    | -1       | 0.7435 | 0.497    | 0.472    | 0.352    | 0.248    | 0.226    | 0.267    | 0.096    | 0.119    | 0.097    | 0.066    |
| 1H71 | D       | E              | 4.6210    | 6.65726   | -1       | 0.4516 | 0.108    | 0.083    | 0.058    | 0.035    | 0.027    | 0.02    | 0.015    | 0.017    | 0.005    |
| 1HHS | A       | D              | 3.716     | 7.5682   | -1       | 0.0836 | 0.412    | 0.312    | 0.254    | 0.263    | 0.225    | 0.232    | 0.115    | 0.115    | 0.094    | 0.068    |
| 1I80 | A       | B              | 3.717     | 7.5682   | -1       | 0.0836 | 0.412    | 0.312    | 0.254    | 0.263    | 0.225    | 0.232    | 0.115    | 0.115    | 0.094    | 0.068    |
| 1QSE | A       | D              | 3.717     | 7.5682   | -1       | 0.0836 | 0.412    | 0.312    | 0.254    | 0.263    | 0.225    | 0.232    | 0.115    | 0.115    | 0.094    | 0.068    |
| 2FTL | A       | B              | 5.893     | 8.9682   | -1       | 0.3708 | 0.362    | 0.218    | 0.16    | 0.102    | 0.032    | 0.046    | 0.018    | 0.01    | 0.021    | 0.021    |
| 1JTD | D       | E              | 4.6210    | 6.65726   | -1       | 0.4516 | 0.108    | 0.083    | 0.058    | 0.035    | 0.027    | 0.02    | 0.015    | 0.017    | 0.005    |
| 1TM1 | A       | B              | 5.0129    | 6.9329    | -1       | 0.7435 | 0.497    | 0.472    | 0.352    | 0.248    | 0.226    | 0.267    | 0.096    | 0.119    | 0.097    | 0.066    |
| 1VLN | A       | B              | 5.0129    | 6.9329    | -1       | 0.7435 | 0.497    | 0.472    | 0.352    | 0.248    | 0.226    | 0.267    | 0.096    | 0.119    | 0.097    | 0.066    |

Table S2: Detailed information, simulation and prediction results of all protein complexes in the benchmark.
The first column of the table is the PDB id of protein complexes in the benchmark.

The second and third column of the table are the chain id of two interacting partners in the protein complexes.

The fourth column of the table, $k_{on}(\text{EXP})$, is the experimentally measured values of association rates for each protein complex in the unit of $M^{-1}s^{-1}$.

The fifth column of the table, $\log(k_{on})$, is the logarithm value of experimental association rates for each protein complex with base 10.

The sixth column of the table gives the assigned class for each protein complex based on the optimal threshold. The value -1 stands for the class which association rates are lower than the threshold and 1 stands for the other class which association rates are higher than the threshold.

The seventh column of the table is the predicted outputs from the neural network. The negative values mean that the protein complexes were predicted as the class which association rates are lower than the threshold, and the positive values mean that the protein complexes were predicted as the class which association rates are higher than the threshold.

The eighth column of the table indicates whether the prediction is true or false by comparing with the experimental value, with 0 means false and 1 means true.

The numbers shown from the nineth to the eighteenth columns of the table are the probabilities of association calculated from the kinetic Monte-Carlo simulations with the distance cutoff values from 16Å to 25Å.