Machine learning based dimensionality reduction facilitates ligand diffusion paths assessment: a case of cytochrome P450cam

J. Rydzewski & W. Nowak

Institute of Physics, Faculty of Physics, Astronomy and Informatics, Nicolaus Copernicus University, Grudziadzka 5, 87-100 Torun, Poland

E-mail: jr@fizyka.umk.pl (JR), wiesiek@fizyka.umk.pl (WN)

Supporting Information

S1. The memetic algorithm.

Optimal and suboptimal diffusion pathways can be reconstructed by minimizing some particular free-energy functional which describes the most important interactions between the protein and ligand. This methodology has been recently proposed by application of memetic algorithms (MA) in searching for the ligand optimal pathways in proteins like the M2 muscarinic receptor, nitrile hydratase and P50cam cytochrome[1]. In Ref. 1, we proposed an algorithm which minimizes the interaction free-energy ΔG between the ligand and protein on-the-fly during MD simulations. Position of the ligand is optimized to minimize ΔG, starting from the initial docking position of the ligand, for each of subsequent ligand intermediates, ending at a state which corresponds to the dissociated ligand-protein complex. The ligand is constraint to the optimal path by the steered MD[2] scheme. The free-energy interaction can be sampled and minimized by many techniques. Out of many scoring-based metaheuristics suitable for minimizing ΔG, we exploited immune algorithm (IA) with additional learning procedures to enhance sampling of ligand positions in a protein. IA is a computational system inspired by the principles of the vertebrate immune system.

Let us consider a set of ligand and protein positions $\boldsymbol{r} \equiv (\boldsymbol{r}_i | i = 0, \ldots, K)$ consisting of the initial docking position of the ligand $\boldsymbol{r}_0$ and ligand intermediates $\boldsymbol{r}_{-0}$, which altogether reconstruct the diffusion path. IA finds the optimal direction of pulling force from $\boldsymbol{r}_{i-1}$ towards the next position $\boldsymbol{r}_i$ by involving an iterative process of mimicking evolution of ith ligand decoys $\boldsymbol{d} \equiv (\boldsymbol{d}_k | k = 1, \ldots, N)$. Each ligand is coded by its position in a protein and orientation. The detailed algorithm for finding the optimal (in terms of the interaction free-energy ΔG) position of ith ligand-protein conformation $\boldsymbol{r}_i$ is described below.

(i) Initially, the N-size population of ligand decoys $\boldsymbol{d}$ is approximated by sampling inside a sphere positioned at the center of mass of a ligand from $\boldsymbol{r}_{i-1}$ of the sampling radius r_s.

(ii) The iteraction free-energy ΔG of the protein and each decoy ligand is calculated by using relation given in the manuscript (Eq. 6).

(iii) The ligand decoys from $\boldsymbol{d}$ with the lowest ΔG have higher probability of being promoted to the population. Selection is performed via roulette scheme.

(iv) The selected decoys are used in the mating and Cauchy deviation procedures[3]. Mating and deviation are performed on randomly picked ligand decoys, according to user-defined probabilities.

(v) An additional local search called hypermutation is applied to all ligand decoys with a defined probability, to decrease ΔG. Random hill mutation climbing (RMHC) was used[4]. In RMHC, a ligand decoy is accepted only if the stochastic Cauchy perturbation leads to a ligand decoy with lower ΔG.

(vi) Steps (ii-v) are repeated until required precision is met, that is, if $|\min \boldsymbol{d}^m - \min \boldsymbol{d}^{m-1}| \leq \delta$, where m is the current iteration. Next, the ith optimal ligand position on the lowest interaction free-energy path is $\boldsymbol{r}_i = \min \boldsymbol{d}^m$. Then, SMD is used to pull $\boldsymbol{r}_{i-1}$ in the direction of $\boldsymbol{r}_i$.

The optimal diffusion path is reconstructed once all the subsequent ligand intermediates $\boldsymbol{r}_{-0}$ are found. The final ligand position must correspond to the state $\boldsymbol{r}_K$ in which the ligand and protein are dissociated. Thus, the ligand diffusion trajectory consists of all the optimal ligand-protein conformations $\boldsymbol{r}$ found by MA.

S2. The list of atom indices from cytochrome P450cam (PDB ID: 2CPP) used to calculate the collective variables

3768, 2963, 2381, 4320, 758, 353, 6446, 6444, 6412, 6443, 6448, 6368

The root-mean-square distance RMSD and the interaction free-energy ΔG mapped on LDCS.

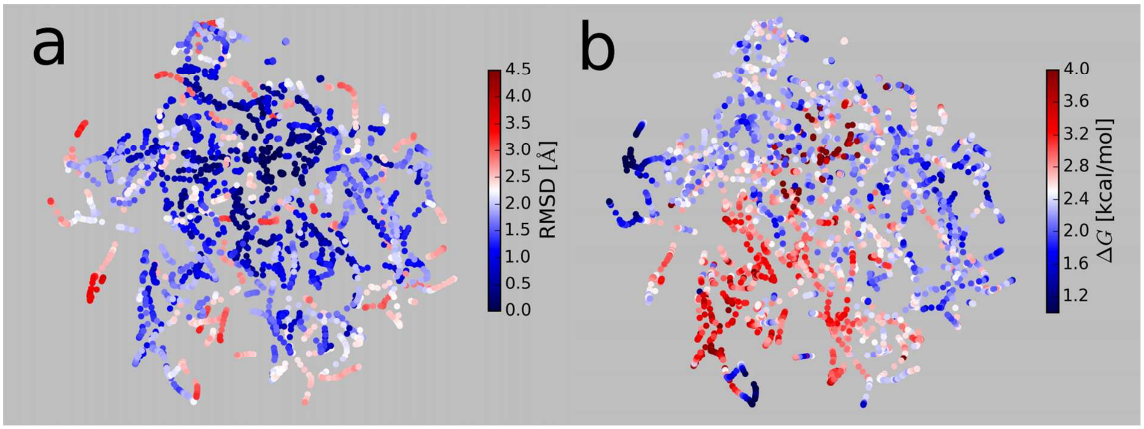

Figure S1: The low-dimensional configuration space of the camphor diffusion from P450cam calculated by t-SNE. (a) The root-mean-square distance RMSD and (b) the interaction free-energy ΔG are mapped instead of **Γ**.

Results of agglomerative clustering

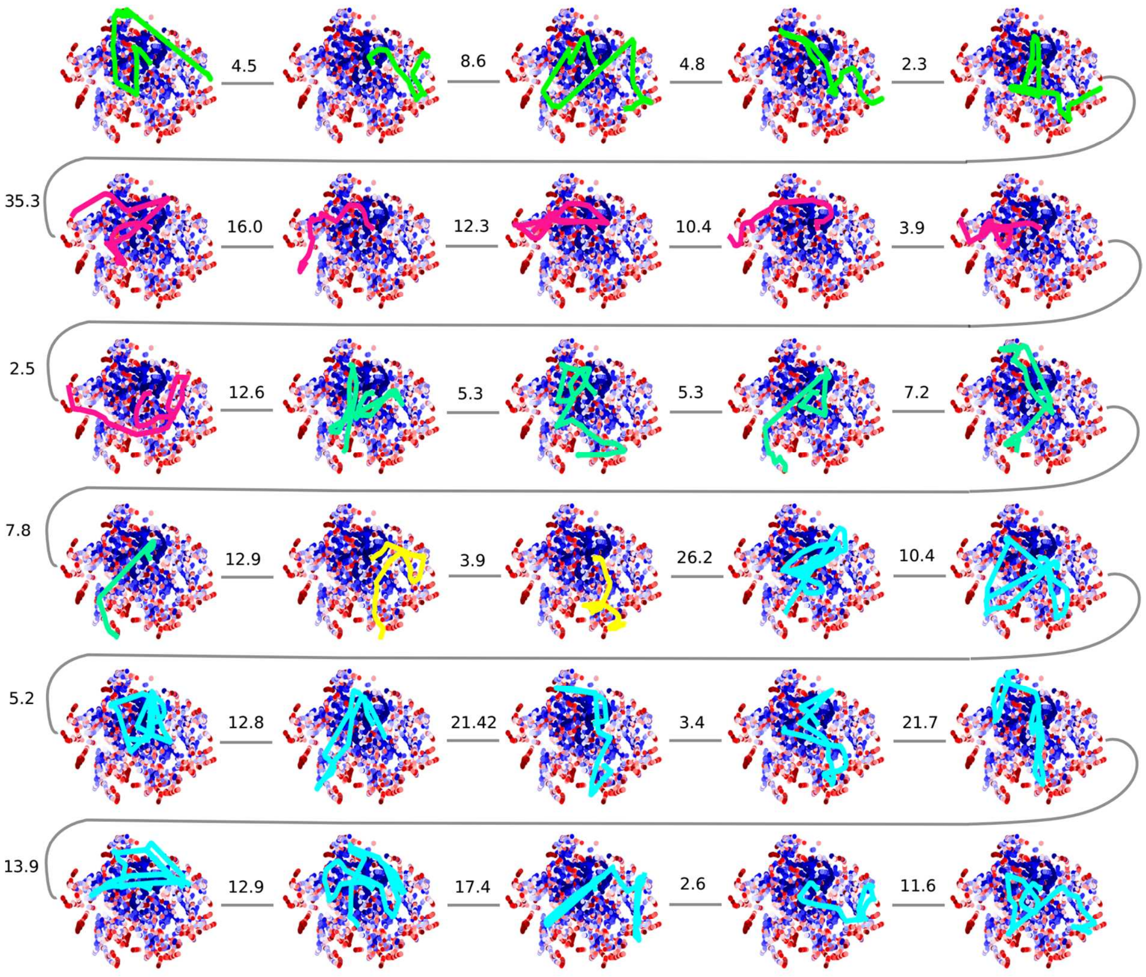

Figure S2: The clusters of ligand diffusion pathways automatically grouped by agglomerative clustering. All 30 enhanced MD trajectories are shown. The Pearson product-moment correlation coefficient r shows total positive correlation ($r = 1$) between the ligand diffusion pathways (PW1-4) observed on graphics and grouped by the agglomerative clustering. Camphor diffusion trajectories were clustered using their topological similarity as calculated by the Fréchet measure (including the last 15 frames from trajectories). Each color corresponds to a particular cluster: PW1 – green, PW2 – pink, PW3 – , PW4 – yellow, interior - turquoise. The edges and numbers between subsequent trajectories depict the Frechet distance between them. Note, that only 15 last structures are taken into account in each trajectory, and to cluster automatically all LDCS trajectories all distances between pairs of trajectories were necessary, not only subsequent, which are shown in this figure.

References

(1) Rydzewski, J.; Nowak, W. Memetic Algorithms for Ligand Expulsion from Protein Cavities. *J. Chem. Phys.* **2015**, *143*, 124101.

(2) Kosztin, D.; Izrailev, S.; Schulten, K. Unbinding of Retinoic Acid from Its Receptor Studied by Steered Molecular Dynamics. *Biophys. J.* **1999**, *76*, 188–197.

(3) Rydzewski, J.; Jakubowski, R.; Nowak, W. Communication: Entropic Measure to Prevent Energy over-Minimization in Molecular Dynamics Simulations. *J. Chem. Phys.* **2015**, *143*, 171103.

(4) Skalak, D. B. Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithms. In *Proceedings of the eleventh international conference on machine learning*; 1994; pp 293–301.