Supporting Information: Quantitative Analysis of Ligand Migration from Transition Networks

Sabyashachi Mishra† and Markus Meuwly†‡*
†Department of Chemistry
University of Basel
Klingelbergstrasse 80, CH-4056, Basel, Switzerland.
‡Chemistry Department
Brown University
Providence, Rhode Island, USA

Clustering of ligand density

Detection of groups of data or clusters sharing important characteristics is an essential step of data analysis. Several approaches exist to detect similarities between data points and to cluster the data, see (1, 2). A widely used strategy to partition data sets is to use the (Euclidean) distance between the points as a measure of their similarity and the distance between the data points and the centroid of the group as the measure of clustering quality. The classification of data set into k clusters is known as k-means clustering. $k-$means clustering is sufficient for the present purpose where we aim to cluster the ligand density into several groups (ligand docking sites). The ligand density is adequately described in terms of three Cartesian coordinates of the ligand with respect to the center of the heme plane which is the zero of the coordinate system. To account for deviations from spherical clusters, fuzzy clustering was used which allows a piece of data to belong to two or more clusters. Here the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster. The "degree of belonging" is related to the inverse of the distance between the data point and the center of the cluster. First, a data point is assigned to the cluster with highest degree of belonging and the centers of the clusters are computed. The

*Address reprint requests and inquiries to M. Meuwly, E-mail: m.meuwly@unibas.ch.

partitioning is then carried out through an iterative process until convergence is achieved, i.e, no more data points change cluster (1, 2).

For a consistent description of ligand density the entire protein was re-oriented so that the least-squares plane containing the four nitrogen atoms of the heme group lies in the xy-plane, with the center of mass of the four nitrogen atoms placed at the origin. The Cartesian distance of the center of mass of NO with respect to the origin is then collected from each snapshot. These data points were then partitioned into clusters, corresponding to Xe1, Xe2, Xe3, Xe4, Xe5, Proximal docking site (PDS), intermediate states between G and H helices (IS1), E and F helices (IS2), and C and F helices (IS3), and the entrance of Channel I (Ch1). Ligand density that escaped the protein into the solvent is grouped into a separate cluster and referred to as WAT. Thus the entire data set was divided into 11 clusters. The clustering method used does not explicitly take into account the conformational changes of the protein, since the partitioning is carried out on the basis of the ligand position in the protein. Nevertheless, the relative position of the ligand is influenced by the protein-ligand interaction and hence, the contributions from protein conformational changes are indirectly included. Furthermore, within the time scale of the simulations (several nanoseconds), only minor changes in the protein conformation are expected. For very long simulations and for problems where protein conformational dynamics plays a key role, such as protein folding problems, the RMSD deviation of protein atoms can also be used for clustering (3).

References

1. Jain, A., and R. Dubes, 1988. Algorithms for clustering data. Prentice Hall.

2. Hastie, R., T. Tibshirani, and J. Fiedman, 2001. The elements of statistical learning. Springer, Berlin.

3. Krivov, S. V., and M. Karplus, 2004. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. U.S.A.* 101:14766–14770.