A comparative analysis of clustering algorithms: O2 migration in truncated hemoglobin I from transition networks.

ID 3402680
Author(s) Cazade, Pierre-Andre; Zheng, Wenwei; Prada-Gracia, Diego; Berezovska, Ganna; Rao, Francesco; Clementi, Cecilia; Meuwly, Markus
Author(s) at UniBasel Meuwly, Markus ;
Year 2015
Title A comparative analysis of clustering algorithms: O2 migration in truncated hemoglobin I from transition networks.
Journal Journal of Chemical Physics
Volume 142
Number 2
Pages / **Article-Number** 025103
The ligand migration network for O2-diffusion in truncated Hemoglobin N is analyzed based on three different clustering schemes. For coordinate-based clustering, the conventional k-means and the kinetics-based Markov Clustering (MCL) methods are employed, whereas the locally scaled diffusion map (LSDMap) method is a collective-variable-based approach. It is found that all three methods agree well in their geometrical definition of the most important docking site, and all experimentally known docking sites are recovered by all three methods. Also, for most of the states, their population coincides quite favourably, whereas the kinetics of and between the states differs. One of the major differences between k-means and MCL clustering on the one hand and LSDMap on the other is that the latter finds one large primary cluster containing the Xe1a, IS1, and ENT states. This is related to the fact that the motion within the state occurs on similar time scales, whereas structurally the state is found to be quite diverse. In agreement with previous explicit atomistic simulations, the Xe3 pocket is found to be a highly dynamical site which points to its potential role as a hub in the network. This is also highlighted in the fact that LSDMap cannot identify this state. First passage time distributions from MCL clusterings using a one- (ligand-position) and two-dimensional (ligand-position and protein-structure) descriptor suggest that ligand- and protein-motions are coupled. The benefits and drawbacks of the three methods are discussed in a comparative fashion and highlight that depending on the questions at hand the best-performing method for a particular data set may differ.
Publisher AIP Publishing
ISSN/ISBN 0021-9606
edoc-URL http://edoc.unibas.ch/41802/
Full Text on edoc Available;
Digital Object Identifier DOI 10.1063/1.4904431
PubMed ID http://www.ncbi.nlm.nih.gov/pubmed/25591387
ISI-Number WOS:000348129700058
Document type (ISI) Article