

Combining MD simulations with sPRE calculations for protein structure prediction

Abstract

MD simulations are a valuable tool to gain insights into protein structure and dynamics. The atomistic resolution paired with the time resolution of pico seconds have enabled looks into the dynamic behavior of protein folding and unfolding. Compared to other methods which yield time-averaged results, MD can be used as a tool to identify sub-states and predict transition paths between such states. However, reaching such conclusions requires large datasets, that can be difficult to obtain. System sizes and their number of atoms scale with r^3 . The folding of some proteins might occur on such long timescales that all-atom simulations might not be feasible. In this manuscript we present a method to combine a limited set of all-atom MD data with a large dataset of coarse-grained MD data and solvent paramagnetic resonance enhancement (sPRE) spectroscopy. We applied this method to three differently ubiquitinated di-ubiquitin chains.

Introduction

Available data

CG data

MD data at coarse-grained resolution was taken from [berg2018towards]. This dataset comprised the ubiquitination sites K6, K29 and K33 where every ubiquitination site contains 10 replicas with 10 ms each. The MARTINI v2.2 force field was used for the non-bonded interaction, whereas the ELNEDIN force field paired with an elastic network iteratively generated with the IDEN method. These forcefields had to be altered to accommodate the isopeptide bond between the proximal subunit's lysine and the distal subunit's C-terminal glycine. The CG data of all 3 ubiquitylation sites encompass approximately 300 ms.

Back-mapped atomistic data

These simulations were used as input-data for a dimensionality reduction algorithm called sketch-map. Here, the high-dimensional phase space of the protein is projected into 2D. The similarity of any two conformations is represented as the distance in sketch-space. Naturally, basins of higher density start to become apparent for statistically significant sub-states of lower free energy. From these basins new all-atom simulations were started from the center of the basin (4 simulations per ubiquitination site with 10 ns each) and from randomly chosen points in the vicinity of the basin (40 simulations per ubiquitination site

with 3 ns each) by using Martini’s backward script paired with some energy minimization to settle the initial high-energy structure.

Extended and rotamer atomistic data

Besides the back-mapped all atom simulations further atomistic data are available from Berg et al. The starting structures of these simulations are created by slightly altering the position between proximal and distal subunit. For every ubiquitination site, there are 12 simulations with 50 ns available. All atomistic simulations use the GROMOS54a7 forcefield with an addition of the isopeptide bond, which was parametrized from regular peptide bonds. The same force field was also used for the rotamer simulations which have been carried out to complement the dataset of atomistic simulations. Here, the different starting structures were created by rotating the χ_3 angle of the ubiquitinated lysine residue in 40 degree steps. The rotamer simulations were intended to accelerate the exploration of the phase space available to the diubiquitin proteins. The data encompass 9 simulations per ubiquitination site with approx 50 ns each. All in all a single ubiquitination site is characterized by 100 ms of CG data and 1.21 ms of AA data.

Methods

Encodermap dimensionality reduction

We used the Encodermap’s auto-encoder neural network to retrieve a dimensionally reduced representation of all aforementioned simulations. The high-dimensional input data was extracted similar to [berg2018towards]. The minimal distances between each residue of the proximal subunit to any residue on the distal subunit (and vice versa) resulted in 152-dimensional input data. The neural network comprised 250, 250, 125, and 2 neurons in a fully-connected layering scheme. Of course, the decoder parts reverses this sequence. Activation functions were set to be tanh for all but the input and bottleneck layers. The sigmoid function which transforms the high-dimensional input space and the low-dimensional latent space was using the following parameters:

Sigma	A	B	sigma	a	b
5.9	12	4	5.9	2	4

The resulting 2D projection was used as input to HDBSCAN clustering.

HDBSCAN clustering

The hierarchical density-based clustering algorithm HDBSCAN has been tried and tested for molecular dynamics systems and especially the low-dimensional

projections of such. After the projections were obtained for every ubiquitination site, the xy-values were fed into HDBSCAN using the ‘leaf’ algorithm, with a minimal cluster size of 2500.

XPLOR calculations

The NMR prediction and refinement program XPLOR NIH in version 3.3 was used to calculate the sPRE and ¹⁵N-relaxation time NMR observables. XPLOR’s new python functionality was used to call XPLOR functions from within python. A script which can be provided a pdb structure file and a psf file was written, which is very similar to run-from-the-mill refinement simulations. The loading, setting-up of potentials is similar to all XPLOR scripts found in online resources. However, our script does not use the IVM and dynamics modules of XPLOR but rather puts out the current observables for the given pdb file and then quits. By doing this in parallel we could retrieve the values for the pSol-potential and the relax-ratio potentials using all all-atom MD simulation frames as inputs within two months.

Normalization of XPLOR values

π