Combining MD simulations with sPRE calculations for protein structure prediciton

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 closer 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 the associated number of atoms scale with r^3 . The folding of some rigid 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 linked di-ubiquitin proteins.

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 is comprised of 10 replicas with 10 ms each. The MARTINI v2.2 force field was used for the non-bonded interaction, whereas bonded interactions have been modeled using the IDEN method. These forcefields had to be altered to accommodate the 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.

Figure: Corase-grained representation of a di-ubiquitin molecule. The backbone is modelled via pink beads, whereas the yellow beads represent the sidechains. In the MARTINI representation every amino acid (except glycine) has only one sidechain bead, but they differ in the parametrization of the interaction potential. On right, the proximal subunit offers its LYS6 residue to be ubiquitinated by the distal unit's GLY76 residue.

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

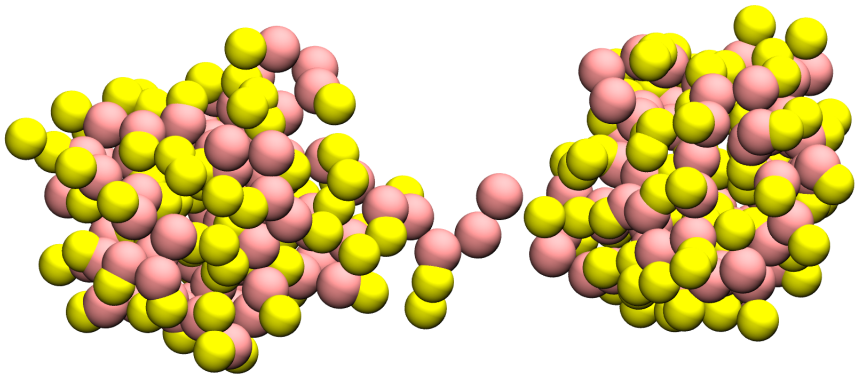

Figure 1: K6_CG

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 sidechain chi3 angle of the ubiquitinated lysine residue in 40 degree steps. The rotamer simulations were intended to accelerate the exploration of the phase pace available to the diubiquitin proteins. The data encompass 9 simulations per ubiquitination site with approximately 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.

Figure: Similar to the above figure, but this time in all-atom resolution using cartoon-representation. Proximal subunit right, distal left. The residues of the isopeptide bond (LYQ and GLQ) are colored orange.

Methods

A general overview over the analysis steps: 1. Extract high-dimensional collective variables 2. Use Encodermap to project into 2D space 3. Use HDBSCAN to extract statistically-significant sub-states of ensemble 4. Use XPLOR NIH to calculate the sPRE values for the all-atom conformations (parallelize this task) 5. Normalize the sPRE data from XPLOR 6. Use sPRE and HDBSCAN to identify some clusters that represent the enxemble available to di-ubiquitin

Extact high-dimensional collective variables

Collective variables represent a type of data, that is somehow aligned with the raw xyz positional data of a molecular dynamics trajectory. Oftentimes CVs are used to visualize complex dynamic processes. One such example might be the simplification of receptor-ligand docking processes by using a receptor-ligand distance as a collective variable. The torsion angles of the backbone are another

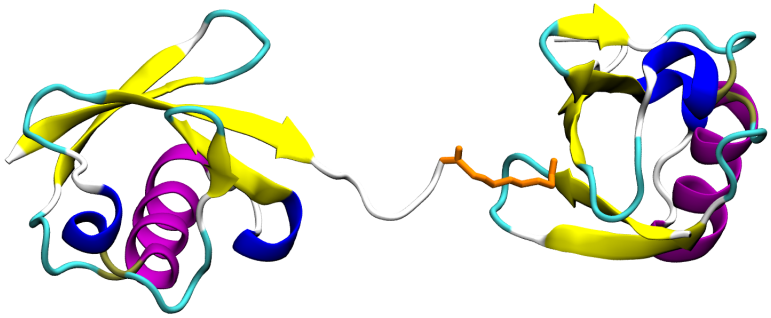

Figure 2: K6_AA

widely used example of collective variables. From the ϕ and ψ dihedral angles a ramachandran plot can be created to visualize and simplify the complex structure of proteins. In this paper we faced the challenge to unify the CG and AA data, meaning CVs needed to be identical for both types of MD data. Torsional angles could not be calculated from CG data, as the backbone N C and CA atoms are unified within one CG bead. Distance matrices between all CA atoms would yield a $\binom{15}{2} = 11476$ dimensional dataset. We chose to use the approcah also used by Berg at al. employing residue-wise minimal distances, which makes the input data 152-dimensional.

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 fed into a dense, fully-connected sequential neural network comprised 250, 250, 125, 2, 125, 250, and 250 neurons. We used the tensorflow python library to minimize three cost functions:

- Center cost: Mean absolute distance between all points and the coordinate origin

$$CenterLoss = \frac{\sum_{i=1}^{n} y_i^2}{n}$$

- Auto cost: Compare the mean absolute difference between input and output

$$AutoLoss = \frac{\sum_{i=1}^{n} \| \sqrt{y_i^2 - \hat{y}_i^2} \|}{n}$$

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 wes using the following parameters:

σ_{highd}	A	B	σ_{lowd}	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-diemnsional 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

$$\pi$$