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 [1]. 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 (Figure 1). 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

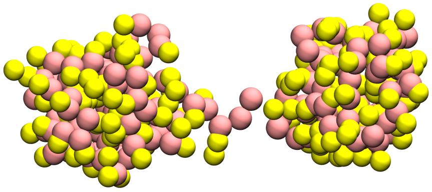

Figure 1: Coarse grained representation of K6-diUbi. Pink beads are unified backbone beads (C_α, N, and C). Yellow beads are unified sidechain beads (one bead for all non-GLY residues). The proximal subunit (right) offers its LYS6 residue to be ubiquitinated by the distal (left) unit's GLY76 residue.

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. As a monomer structure the 1UBQ crystal structure from the protein database was used [2]. 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.

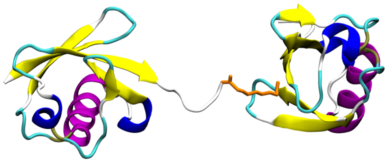

Figure 2: All-atom representation of K6-diUbi. The structure in this figure was used as a basis for Figure 1 and the location of proximal and distal units are identical. Secondary structure has been visualized as cartoon representation and colored accordingly. The residues of the isopeptide bons (GLY76 (left) and LYS6 (right)) 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 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 [3]. 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

$$CenterLoss = \frac{\sum_{i=1}^{n} \hat{y}_i^2}{n}$$

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

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

- Encodermap cost: Compare the input and the latent space by transforming it with a sigmoid function.

$$EncoderMapLoss = \frac{1}{m} \sum_{i \neq j} \left[SIG_h(R_{ij}) - SIG_l(r_{ij}) \right]^2$$

,

where SIG is a sigmoid function originally devised by Ceriotti et al. for their sketch-map dimensionality reduction algorithm [4].

$$SIG_{\sigma,a,b}(r) = 1 - \left(1 + \left(2^{\frac{a}{b}} - 1\right) \left(\frac{r}{\sigma}\right)^a\right)^{-\frac{b}{a}}$$

R_{ij} and r_{ij} are the pairwise distance between all points in the high-dimensional input space and the 2-dimensional latent space. They are defined using the pairwise distance matrices $boldsymbolR_{i,j}$:

$$\boldsymbol{R}_{i,j} = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\ r_{2,1} & r_{2,2} & & \\ \vdots & & \ddots & \\ r_{n,1} & & & r_{n,n} \end{pmatrix}$$

without the diagonal elements (which are always 0) and vectorizing them:

$$R_{ij} = \mathrm{vec}(\boldsymbol{R}_{i,j,i\neq j}) = \begin{pmatrix} r_{1,2} \\ r_{1,3} \\ \vdots \\ r_{1,n} \\ r_{2,1} \\ \vdots \\ r_{n,n-1} \end{pmatrix}$$

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 [5]. After the 2D 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, resulting in 13, 18, and 24 selected clusters, for K6, K29, and K33, respectively. The clusters are enumbered based on the number of points/conformations in them, with cluster 0 being the largest cluster and so on. Another point to note is, that HDBSCAN offers a great advantage by distinguishing between clusters and noise, which - speaking from an ensemble viewpoint - is comprised of unique structures that carry no significance.

Look at a specific cluster

As an example cluster 0 of K6 diUbi will be further investigated. This cluster consists of 131969 conformations sampled by XXX independent CG and XXX independent AA simulations and thus makes up roughly 12% of the complete K6 diUbi ensemble. Only 3% of the conformations in this cluster are from all-atom simulations (Figure 5). Visualizing 10 of the approximately 4000 all-atom conformations shows, that they are indeed structurally similar and this task succeeded (Figure 6).

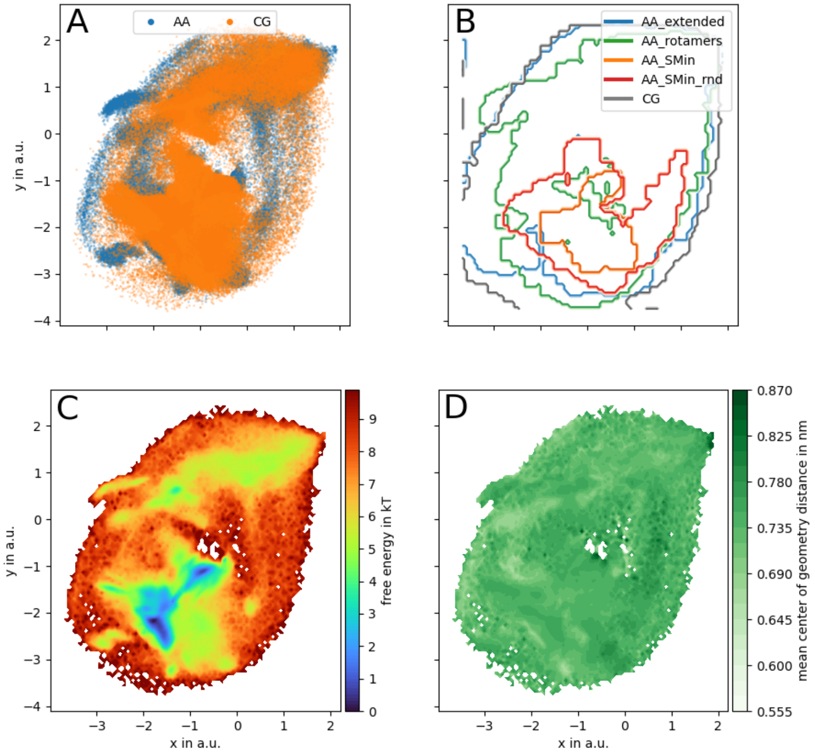

Figure 3: 2D projections of all CG and AA conformations of K6 diUbi. When the low-dimensional coordinates are plotted as a scatter plot, every point can be associated with one conformation sampled by a MD simulation (A). Orange points belong to CG simulations, blue points to AA. The closeness of two points. In (B) the areas occupied by specific MD simulation trajectories is visualized. Here, it can be seen, that the long AA extended simulations sampled more unique conformations, when compared to the limited regions of the sketch-map minima and sketch-map minima random sims. It should be noted, that the AA_rotamers were a more sucessfull strategy in creating unique conformations, when compared to SMin and SMIN_rnd. However, the AA_rotamer simulations are inherently starting-structure biased and might not yield a representative i.e. biased ensemble. That's why the long (and with than more unbiased) CG simulations are used to obtain a more correctly weighted ensemble. The Encodermap projections correlate to many other structural properties of the system. So can the distance between the center of geometries of the Ub subunits be used to color the projection and distinct regions can be observed (D).

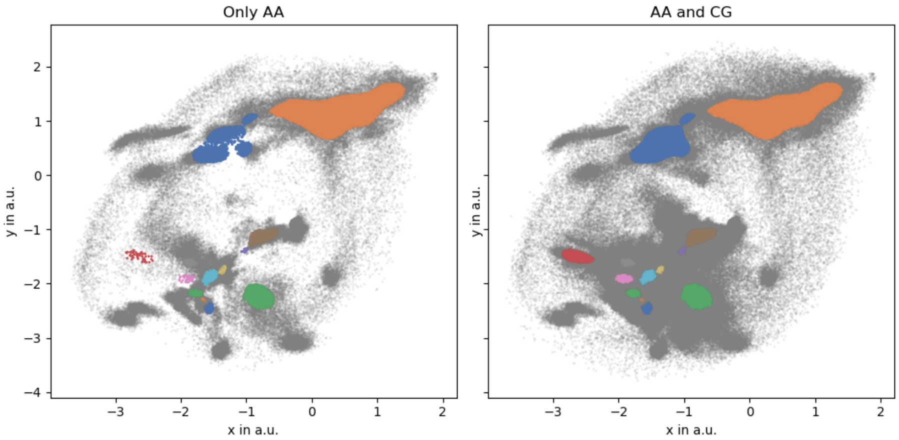

Figure 4: The Encodermap projection of K6 colored according to cluster membership. The advantage of using combined CG and AA data becomes apparent, when the red cluster is investigated. In a combined dataset this cluster contains more than 2500 structurally similar conformations in a dense basin (i.e. statistically significant sub-state of the K6 diUbi system), but multiple AA simulations were not able to sample these structures sufficiently, which would have removed these structures from the list of candidates for further analysis.

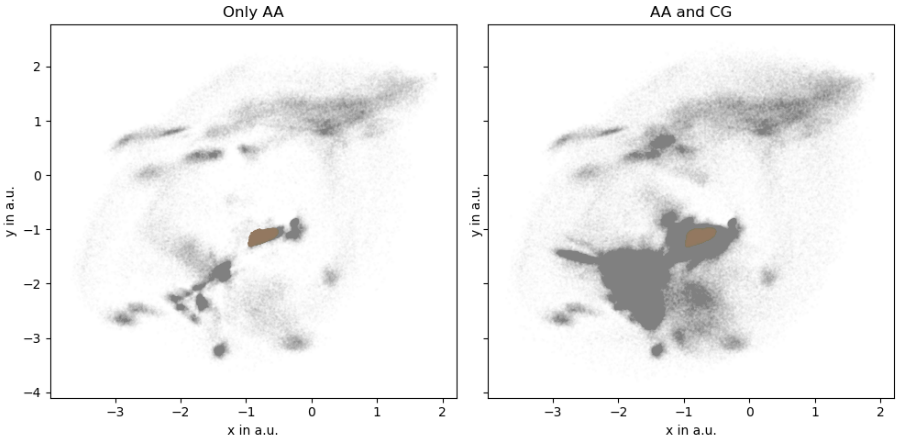

Figure 5: Encodermap projections of K6 diUbi, with cluster 0 highlighted.

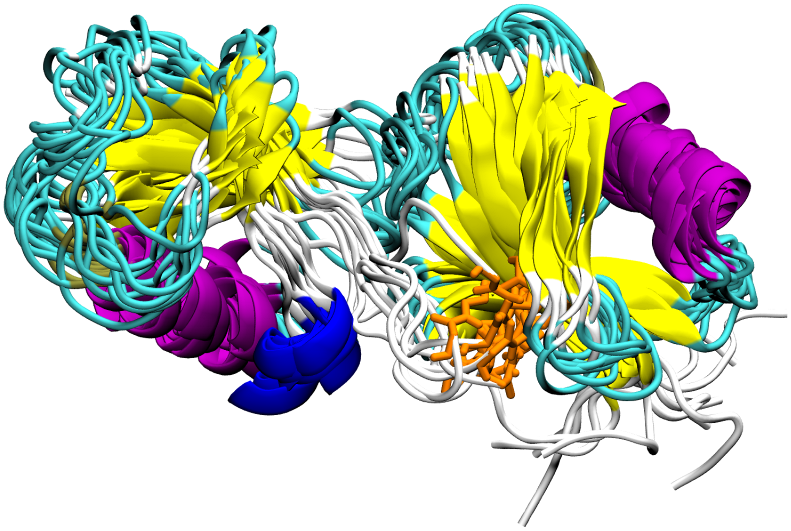

Figure 6: 10 conformations from cluster 0 of K6 diUbi. The structures are rendered as cartoon representation and colored accordingly. The isopeptide LYS and GLY residues are visualized as orange licorice (proximal subunit on the right, distal subunit on the left). It is notable, that the structural difference in the residues taking part in the isopeptide bond is larger, when compared to the protein backbone. The α-helices seem to overlap more precisely.

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

42

[1] A. Berg, O. Kukharenko, M. Scheffner, and C. Peter, "Towards a molecular basis of ubiquitin signaling: A dual-scale simulation study of ubiquitin dimers," *PLoS computational biology*, vol. 14, no. 11, p. e1006589, 2018.

[2] S. Vijay-Kumar, C. E. Bugg, and W. J. Cook, "Structure of ubiquitin refined at 1.8 Åresolution," *Journal of molecular biology*, vol. 194, no. 3, pp. 531–544, 1987.

[3] T. Lemke and C. Peter, "EncoderMap: Dimensionality reduction and generation of molecule conformations," *Journal of chemical theory and computation*, vol. 15, no. 2, pp. 1209–1215, 2019.

[4] M. Ceriotti, G. A. Tribello, and M. Parrinello, "Simplifying the representation of complex free-energy landscapes using sketch-map," *Proceedings of the National Academy of Sciences*, vol. 108, no. 32, pp. 13023–13028, 2011.

[5] L. McInnes, J. Healy, and S. Astels, "Hdbscan: Hierarchical density based clustering," *Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.