De novo ensemble modeling suggests that AP2-binding to disordered regions can increase steric volume of Epsin but not Eps15

N. Suhas Jagannathan
Duke-NUS Medical School, National University of Singapore, 8 College Road, Singapore 169857
Email: suhas@duke-nus.edu.sg

Christopher W. V. Hogue
600 Epic Way Unit 345, San Jose CA 95134
Email: chogue@blueprint.org

Lisa Tucker-Kellogg
Cancer & Stem Cell Biology, and Centre for Computational Biology
Duke-NUS Medical School, 8 College Road, Singapore 169857
Email: Lisa.Tucker-Kellogg@duke-nus.edu.sg

Proteins with intrinsically disordered regions (IDRs) have large hydrodynamic radii, compared with globular proteins of equivalent weight. Recent experiments showed that IDRs with large radii can create steric pressure to drive membrane curvature during Clathrin-mediated endocytosis (CME). Epsin and Eps15 are two CME proteins with IDRs that contain multiple motifs for binding the adaptor protein AP2, but the impact of AP2-binding on these IDRs is unknown. Some IDRs acquire binding-induced function by forming a folded quaternary structure, but we hypothesize that the IDRs of Epsin and/or Eps15 acquire binding-induced function by increasing their steric volume. We explore this hypothesis *in silico* by generating conformational ensembles of the IDRs of Epsin (4 million structures) or Eps15 (3 million structures), then estimating the impact of AP2-binding on Radius of Gyration (R_G). Results show that the ensemble of Epsin IDR conformations that accommodate AP2 binding has a right-shifted distribution of R_G (larger radii) than the unbound Epsin ensemble. In contrast, the ensemble of Eps15 IDR conformations has comparable R_G distribution between AP2-bound and unbound. We speculate that AP2 triggers the Epsin IDR to function through binding-induced-expansion, which could increase steric pressure and membrane bending during CME.

Keywords: Intrinsically disordered proteins; Clathrin-mediated endocytosis; coarse-grained methods; all-atom ensemble modelling; steric crowding; molecular crowding; membrane biophysics.

1. Introduction

Protein science in the twentieth century assumed as dogma the direct causality from protein sequence to structure, and from structure to function. This structure-to-function link has increasingly come into question with the discovery that many proteins have intrinsically disordered regions (IDRs),[1,2] characterized by the presence of long contiguous amino acid sequences (> 30 residues) with no discernible secondary structure. IDRs have been found in more than 40% of the eukaryotic proteome and are distributed across multiple pathways[3–9].

The surprising evolutionary conservation of many IDRs from simple to higher eukaryotes suggests that IDRs have biological functions, beyond serving as linkers or filler-regions between functional domains[10]. Experimental techniques to probe IDPs (e.g., NMR and SAXS[11]) are limited due to the lack of stable folding. To complement this limitation, computational methods for studying IDRs have been burgeoning, representing IDRs as statistical ensembles rather than as stably-folded structures[12]. Computationally-derived ensembles have been constructed using different methods (e.g., Molecular dynamics, Monte Carlo methods[13–15]) based on experimentally-determined constraints or *de novo* assumptions[16,17]. Atomic force fields for disordered regions have also improved the accuracy of IDP representations *in silico*[18,19]. The use of both experimental and theoretical approaches has generated important strides for the field.

Toward the elucidation of IDR function, multiple examples have been catalogued of binding-induced transitions of disordered regions into ordered structures, where the ordered structure has a biological function. IDRs have also been found to function without folding into an ordered structure, by utilizing molecular crowding to induce steric pressure. In essence, disordered regions can function through their entropic forces[20]. In an important 2015 study[21], natural and engineered IDRs were capable of bending a cell membrane, altering organelle morphology, and regulating the cargo capacity of clathrin-mediated endocytosis (CME).

Clathrin-mediated Endocytosis (CME) involves encapsulation and internalization of extracellular cargo in a Clathrin-coated vesicle. CME depends on the principal protein Clathrin and on accessory proteins such as Epsin, Eps15, Adaptor proteins (APs), Auxilin, Amphiphysin, Dynamin, Intersectin, Synaptojanin, and Synaptotagmin[22]. Although each protein has been characterized individually, much remains unclear about the orchestration of these proteins and the mechanistic underpinnings of CME. Most CME proteins contain binding/interaction sites for multiple other CME proteins, and many of the proteins are heavily enriched for IDRs[9]. The interplay between binding, IDRs, and steric pressure is unclear. In particular, prior work showed that IDRs can exert steric pressure for membrane bending in CME due to molecular crowding of the unfolded state, but the impact of binding on the unfolded state is unknown. We hypothesize that binding can induce expansion rather than folding of an IDR conformational ensemble, leading to increased steric volume. If correct, this could be crucial for regulating steric pressure during morphogenesis of the CME vesicle.

To study how binding affects sterics, we focus on the IDRs of Epsin and Eps15, which are long regions harboring multiple instances of sequence motifs for binding the alpha-subunit of the CME adaptor protein AP2 (AP2α). Epsin is a ~600 aa protein (with isoforms ranging from 576-640 aa) that contributes to membrane bending in CME[21,23]. Epsin contains an N-terminal region[23,24] that binds the membrane, followed by an IDR (~400 aa) until the C terminus. This IDR contains a Clathrin/AP2 binding (CLAP) region with two binding sites for Clathrin and 8 instances of the motif DPW for binding AP2α [9]. The sequence downstream of the CLAP region contains binding motifs for other CME proteins such as Eps15 and Intersectin. Eps15 is an 896 aa protein that localizes to the rim of the Clathrin coat in growing vesicles[25] suggesting that it might contribute to membrane bending. Eps15 contains an N-terminal region for binding Epsin and other CME proteins, and a C-terminal IDR (~350 aa) that contains 15 instances of the sequence motif DPF for binding to AP2α [26,27]. In summary, both Epsin and Eps15 IDRs are candidates for regulating the steric pressure of

CME membrane bending in ways that might be regulated by AP2-binding. Computational modeling is needed for prioritizing which protein(s) to test in biophysical experiments.

In this study, we generate all-atom *de novo* conformational ensembles for the disordered regions of Epsin and Eps15 using the FoldTraj program[13,14] of the TraDES package, and we investigate whether the IDR conformations that can accommodate binding of multiple copies of the AP2α domain have larger conformational volumes. Results show that conformations of the Epsin IDRs that allow binding of more copies of AP2α possess longer lengths and larger radii of gyration (R_G), indicating larger steric volumes for the IDR. This finding is non-trivial and does not hold true for Eps15. Binding-induced-expansion of Epsin might serve to exert steric pressure on the plasma membrane for regulating CME curvature or cargo.

2. Methods

2.1. *Generation of structural ensembles*

Protein sequences of human Epsin Isoform 2 (Uniprot ID: Q9Y6I3-1) and Human Eps15 Isoform 1 (Uniprot ID: P42566) were analyzed using the disorder predictor IUpred to identify intrinsically disordered regions (IDRs). All known AP2 binding motifs were located within the identified IDRs. Residues 232-471 of Epsin and residues 498-830 of Eps15 covered the IDRs plus an adjacent helix from the ordered region (N-terminal to the IDR, for alignment purposes). The regions were represented using all-atom models and input into FoldTraj[13] for generating conformational ensembles. During ensemble generation, the multiple repeated binding motifs (DPW for Epsin, and DPF for Eps15) were constrained to adopt fixed phi and psi angles taken from the observed phi and psi angles of the PDB structures of AP2α bound to either a small Epsin DPW motif (RCSB ID: 1KY6) or a small Eps15 DPF motif (RCSB ID: 1KYF). No other structural constraints were imposed, and the FoldTraj program was run until we obtained 4 million feasible structures for the Epsin IDR and 3 million feasible structures for the Eps15 IDR.

2.2. *Filtering Epsin conformers to mimic the effect of Plasma membrane*

Since the N-terminal domain of Epsin is attached to the membrane, the membrane is a boundary that the generated structural ensemble can never be allowed to traverse, to mimic biological constraints. This boundary constraint was imposed by first aligning all 4 million generated Epsin IDR structures by their N-terminal helices (from the ordered region), and then discarding structures that intersected the plane, using a plane 15 Å above the helix of the disordered structures, where 15 Å is the maximum diameter of the Epsin N-terminal ENTH domain (computed using RCSB ID: 1H0A).

2.3. *Docking AP2α to the IDRs by superposition*

The ensemble structures were then subjected to docking-by-superposition to dock AP2α domain to each of the binding motifs individually. This was achieved by superimposing the motifs from the generated Epsin or Eps15 structures (previously constrained to resemble bound peptides) and the corresponding peptide bound to the AP2α domain (from PDB), using the *salign* module in the

TraDES package (http://trades.blueprint.org/). AP2α was rigidly docked to each of the IDR conformers, at each of the 8 DPW motifs for Epsin and each of the 15 DPF motifs of Eps15, thus generating a total of ~15 million dockings for Epsin and ~45 million dockings for Eps15, where each docking contains an IDR with one copy of AP2α bound at one site. For each IDR conformer, we then used the set of singly-docked structures to perform docking at other sites and catalogue the number of simultaneous binding sites at which AP2α docking would be feasible (i.e., when binding multiple copies of the AP2α domain). Because flexible docking is prohibitive to compute for tens of millions of structures, binding feasibility was estimated by assuming that flexibility could compensate for a limited numbers of collisions (a limited number of intermolecular clashes during rigid superposition). For each "bound" structure, we computed the number of inter-chain Van derWaals clashes using the *Crashchk* module of TraDES. After viewing the distributions of collisions over the entire ensemble, a threshold value of 200 intermolecular collisions was chosen as the unacceptable level for a docking to be discarded as infeasible. IDR conformers that could "dock" with AP2α having fewer collisions than this threshold were considered feasible.

2.4. *Computation of Protein energies*

To approximate the energies of the structures in the generated ensembles for Epsin and Eps15 IDRs, we used the program FoldX[28], which calculates the contributions of 22 energy terms to the stability of a protein. Some of these parameters were developed for globular proteins, but since disordered regions do not behave like folded proteins, and since the structures generated by FoldTraj were not subjected to energy minimization, we used only the following subset of energy terms from FoldX: hydrogen bonds, electrostatic interactions, polar and non-polar solvation energies, torsional clashes, and weak penalization for VdW clashes.

3. Results

3.1. *Epsin and Eps15 C terminal regions are predicted to have large disordered C-terminal domains*

Prior characterizations of Epsin[29] and Eps15[30] have suggested the presence of intrinsically disordered regions near the C-termini. Consistent with this literature, sequence analysis of human Epsin (Uniprot: Q9Y6I3-1) and Eps15 (Uniprot: P42566) using IUPred also predicts C-terminal regions with high propensities for disorder (Fig. 1A-B). Compositional analysis of the amino acids in these disordered regions revealed high levels of Proline and charged residues, as expected in IDRs (data not shown). In Epsin, the IDR predicted by IUPred extends from residue 253 to the C-terminus (residue 662), and is predicted to be completely disordered. It contains 8 instances of the putative AP2α binding motif DPW. Since the last DPW motif occurs at position 463, we define the region from residues 232-471 as the disordered AP2-binding region of Epsin (hereafter Epsin-IDR). In contrast, Eps15 is an 896 aa protein that was predicted to have a mostly disordered region (with small pockets of ordered residues) starting from position 350. This disordered region contains 15 instances of the motif DPF believed to bind AP2α. Since the last motif occurs at site 825, we define the region 498-830 to represent Eps15-IDR for subsequent analyses. For both Epsin and Eps15, the

chosen regions also include a predicted helix, N-terminal to the IDR, to provide a reference frame for subsequent alignment of the generated conformers. Fig. 1A-B also show secondary-structure predictions for Epsin and Eps15 given by JPred4[31]. Note the absence of secondary structure in the Epsin-IDR and Eps15-IDR, except for the N-terminal helices.

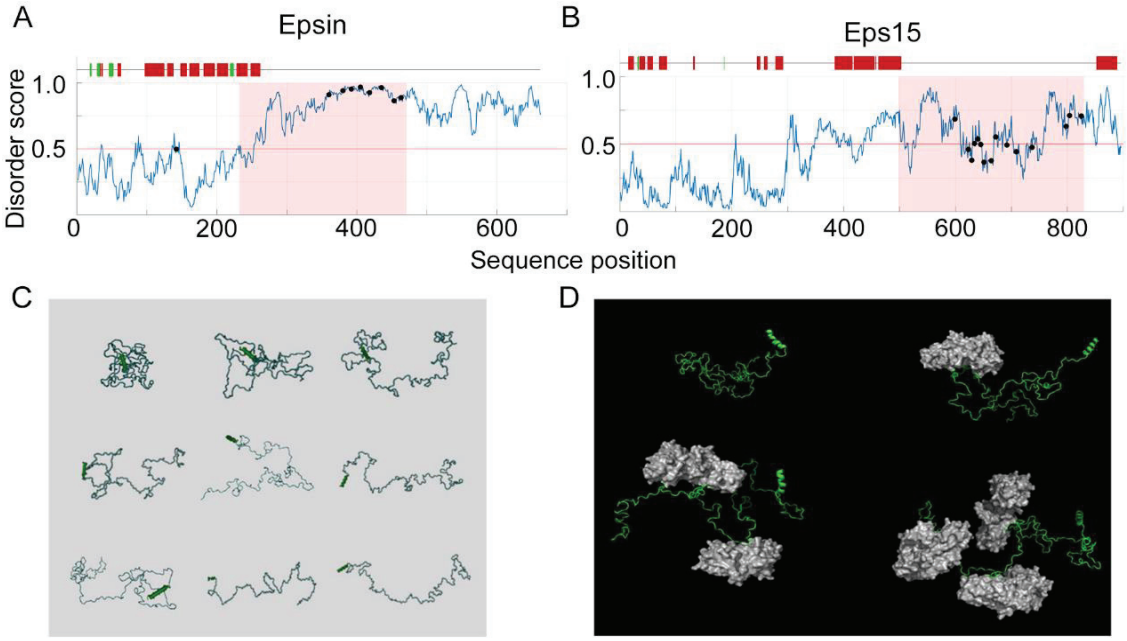

Fig. 1. Generation of conformational ensembles of the IDRs of Epsin and Eps15. (A-B) IUPred disorder prediction scores for each residue of Epsin (A, Uniprot ID: Q9Y6I3-1) and Eps15 (B, Uniprot ID: P42566). Residues with disorder scores above 0.5 (horizontal red line) are considered disordered. Shown above the plots are secondary structure predictions using JPred4 (Red = helix; green = sheet; black line = no SS). Shaded window within the plot region represents the portion of the sequence considered for conformational ensemble generation for Epsin (232-471) and Eps15 (498-830). Black markers within the plot show start of the sequence motifs that bind to AP2α. (C) Example structures of the Epsin-IDR generated using the FoldTraj program, which generates sterically-feasible structures. Note that the set of feasible structures has a wide range of hydrodynamic radii. (D) Example structures of different Epsin-IDR conformers docked to 0, 1, 2 or 3 AP2α domains

3.2. *Generating the conformational ensemble for Epsin-IDR and Eps15-IDR*

To generate a conformational ensemble for both Epsin-IDR and Eps15-IDR, we used the FoldTraj program of the TraDES suite[13]. FoldTraj generates individual conformers by performing a random walk through allowed Ramachandran-space for each amino acid in the sequence (random draw of allowed torsion angles with replacement), while satisfying user-provided constraints of secondary structure (if any). Since FoldTraj uses backtracking to avoid potential steric clashes, the generated conformers are sterically-feasible all-atom structures of the disordered regions. Although FoldTraj is a less rigorous tool than MD to generate conformational ensembles, the characteristics of its output ensembles have shown agreement with SAXS data[32]. Using this program we generated an ensemble of 3 million conformers for Eps15-IDR. Since Epsin is bound to the membrane (unlike Eps15), the presence of the bilayer imposes additional constraints on the available space for Epsin-IDR. Hence, we generated more conformers for Epsin than Eps15 (~4 million) and followed this with a filtering process in which a plane mimicking the membrane bilayer served to remove

conformers that traversed it. We will hereafter refer to the 3 million Eps15 structures and the plane-filtered Epsin-IDR structures as the initial ensembles for the respective proteins. Fig 1C shows examples of Epsin-IDR conformers generated using FoldTraj.

3.3. *Subsets of the Epsin-IDR ensemble that allow binding to more copies of AP2α show increased size metrics*

To each of the structures in the Epsin-IDR initial ensemble, we docked AP2α by superposition (see Methods) to each DPW binding motif in the IDR, and estimated the feasibility of each docking based on the number of interchain atom-collisions. This results in subsets of the initial ensembles, categorized by the number of occupied binding sites (example in Fig 1D), and the exact sites occupied. We then compared the bound subsets against the initial ensemble using two metrics of size: the end-to-end distance (EED), and the radius of gyration (R_G). End-to-end distance refers to the distance between the first and last Cα of the IDR. Radius of gyration (R_G) is defined as the sum of root-mean-squared distances from each of the atoms to the protein center of mass. The R_G describes the relative compactness of a protein structure. For a given sequence length, folded proteins have low R_G, while extended proteins have high R_G. The R_G distribution of the original unfiltered ensemble of 4 million Epsin structures was normally distributed around 43.17 ± 9.72 Å. For comparison, globular proteins of size 200-300 aa typically assume conformations with R_G around 20 Å. The unconstrained nature of FoldTraj conformations result in a predominantly extended ensemble, but it also includes compact structures. Table 1 show that subsets of the Epsin-IDRs that allowed binding to more copies of AP2α showed increased size, as measured by both EED and R_G, compared with subsets that did not permit binding. Each successive AP2α binding resulted in an increase of ~7-8 Å in EED and ~3 Å in R_G.

Table 1. Statistics of end-to-end distance (EED) and radius of gyration (R_G) for the Epsin-IDR ensembles that allow binding to increasing numbers of AP2α domains.

Screening criteria	Num. Stuctures	End-to-end distance (EED)			Radius of Gyration (R_G)		
		Mean	Median	Std. Dev	Mean	Median	Std. Dev
ORIGINAL ENSEMBLE	3932805	101.85	99.19	39.81	43.17	41.97	9.72
INITIAL (PLANE FILTERED)	1867311	110.06	107.94	40.37	43.12	41.88	9.86
SINGLE AP2α	463709	114.53	112.60	41.05	44.53	43.39	10.07
DOUBLE AP2α	62682	122.97	121.45	41.91	47.20	46.21	10.27
TRIPLE AP2α	8083	130.68	129.50	42.59	49.64	48.73	10.39
QUADRUPLE AP2α	824	138.71	138.09	43.22	52.16	51.32	10.48

Fig 2A and 2B show examples of the distribution of EED and R_G (respectively) in the initial ensembles of Epsin-IDR compared with a representative subset of Epsin-IDRs that allowed AP2α binding at 4 sites (sites 1, 2, 3, and 7) simultaneously.

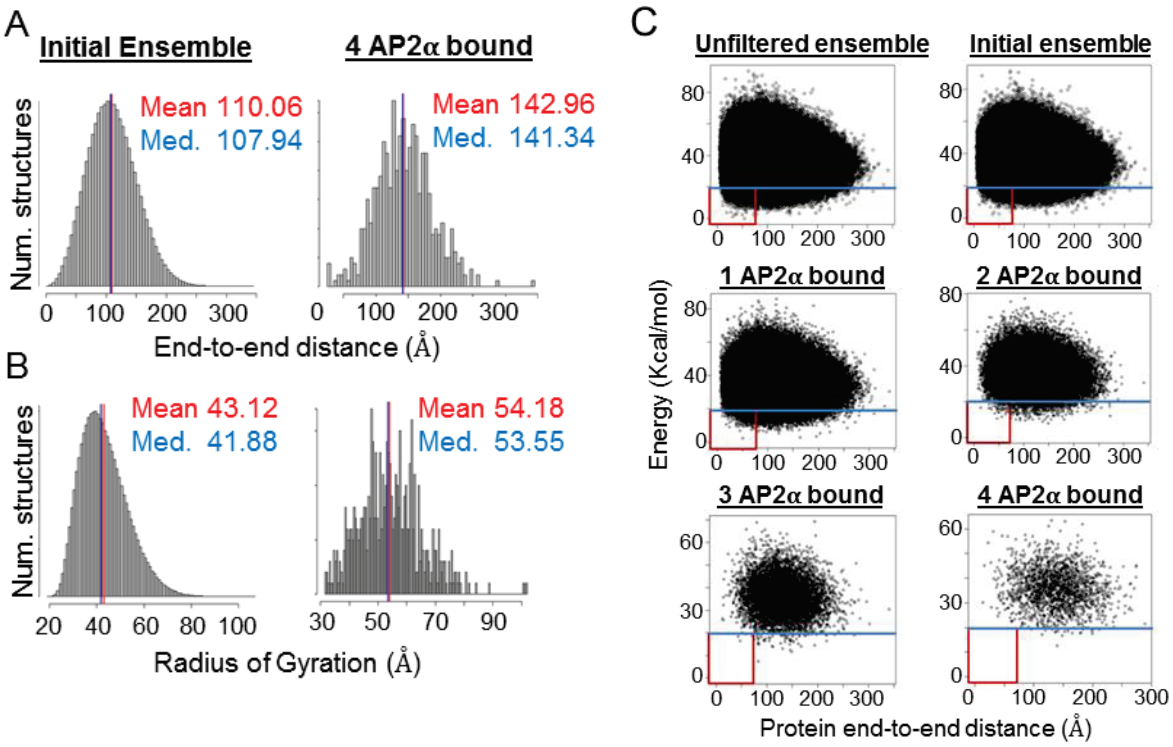

Fig. 2. AP2α binding to the Epsin-IDR causes expansion of EED and disproportionate depopulation of compact low-energy structures. (A) Example distribution of end-to-end distances or (B) radius of gyration in the ensemble of Epsin-IDR structures, with either 0 (*left*) or a representative ensemble of Epsin-IDR with 4 (*right*) AP2α domains bound (AP2 bound at sites 1,2,3 and 7). AP2α binding can be seen to cause an increase in both metrics. Complete statistics can be found in Table 1 (C) Distribution of energies of the ensemble of Epsin-IDRs computed using FoldX plotted against end-to-end distance. With more AP2α domains bound, there is preferential depopulation of the small-length low energy structures (red box) leading the ensemble to shift towards even longer lengths than expected with panels A and B.

3.4. *Epsin ensembles that allow binding of more copies of AP2α exhibit preferential depopulation of compact low-energy structures*

The impact of AP2-binding on the size and energy of the Epsin-IDR feasible ensemble is shown in Fig 2C, which plots the FoldX-computed energy versus the end-to-end distance for each of the Epsin-IDR structures (see Methods). The subsets that allow binding to more copies of AP2α clearly contain fewer conformers than the unbound ensemble, which is a trivial consequence of having more constraints. The non-trivial question is whether the bound subsets are uniformly depopulated across different energy levels and different conformational sizes. In particular, we must test whether compact structures become infeasible or energetically unfavorable upon binding, in greater proportion than other structures. We define compact structures as those with less than a quarter of the maximum EED (defined as 25% of the 99th percentile for EED in the original ensemble). Similarly, we define low-energy structures as those having less than 25% of the 99th percentile of FoldX-computed energy. In Figure 2C, the threshold for low-energy is marked by blue lines, and

the combined threshold for compact and low energy is marked by a red box at the bottom-left corner of each frame. Comparing across the frames from unbound to bound and multiple-bound, the red box of compact low-energy structures becomes disproportionately depopulated. In other words, AP2 binding causes the overall ensemble to shift toward the upper-right. We quantified this trend by counting the number of conformations above and below the EED threshold, and above and below the energy threshold (four regions of the state space); Table 2 lists the totals according to the number of AP2α domains bound. Results show that with more AP2α binding, there is an increase in the fraction of high-EED, high-energy structures at the expense of the other three quadrants, especially the compact structures. Binding each copy of AP2α causes a slight decrease (5-12%) in the high-EED low-energy populations, a 25-26% depopulation of the compact high-energy populations, and a 26-55% decrease in the compact low-energy populations (going almost to 0% with 4 copies of AP2α bound). Low energy structures remain feasible with high-EED, but compact structures of all types become increasingly infeasible. This indicates that AP2α binding causes conformational expansion and preferential depopulation of the compact low energy structures.

Table 2. Relative proportions of Epsin-IDR conformers in different regions of the Energy – EED (end-to-end distance) state space. Thresholds for EED: 75 Å; Energy: 20 Kcal/mol

SCREENING CRITERIA	Low EED Low Energy	Low EED High Energy	High EED High Energy	High EED Low Energy
SINGLE AP2α	0.23	17.22	81.37	1.18
DOUBLE AP2α	0.15	12.62	86.11	1.12
TRIPLE AP2α	0.11	9.34	89.5	1.05
QUADRUPLE AP2α	0.05	6.88	92.15	0.92

3.5. *Subsets of the Eps15-IDR ensemble that allow binding of more copies of AP2α show no change in size metrics or preferential depopulation by energy or size*

Similar to the Epsin analysis, we docked AP2α individually to each of the 15 DPF binding sites in every structure to obtain subset ensembles of Eps15 IDR structures that allowed different stoichiometries and positions of AP2α binding. In general, the dimensions of the Eps15 ensembles were greater than that of the Epsin-IDR ensembles, but this could be attributed to the difference in sequence lengths of the two IDRs (Epsin-IDR: 240aa; Eps15-IDR: 333aa). More importantly, unlike Epsin, ensembles of Eps15-IDR that allow binding of more copies of AP2α did not show a significant increase in either EED or R_G (Table 3). Energy-EED thresholding also do not show preferential enrichment or depopulation of any one quadrant (Table 4). These results also serve to illustrate that Epsin's size-increase and preferential depopulation of compact structures are not a trivial property of all IDRs, nor an artefact of the docking analysis.

Table 3. Statistics of end-to-end distance (EED) and radius of gyration (R_G) for the Eps15-IDR ensembles that allow binding to increasing numbers of AP2α domains.

Screening criteria	Num. Stuctures	End-to-end distance (EED)			Radius of Gyration (R_G)		
		Mean	Median	Std. Dev	Mean	Median	Std. Dev
INITIAL ENSEMBLE	3000000	124.84	121.11	50.80	53.80	52.30	12.20
SINGLE AP2α	1238955	128.05	124.53	51.55	54.97	53.52	12.42
DOUBLE AP2α	509819	130.60	127.21	52.18	55.89	54.47	12.63
TRIPLE AP2α	220656	131.57	128.16	52.54	56.21	54.79	12.80
QUADRUPLE AP2α	106947	130.66	127.07	52.51	55.85	54.36	12.84

Table 4. Relative proportions of Eps15-IDR conformers in different regions of the Energy – EED (end-to-end distance) state space. Thresholds for EED: 75Å; Energy: 80 Kcal/mol

SCREENING CRITERIA	Low EED Low Energy	Low EED High Energy	High EED High Energy	High EED Low Energy
SINGLE AP2α	0.17	15.53	82.98	1.32
DOUBLE AP2α	0.17	14.59	83.85	1.39
TRIPLE AP2α	0.17	14.31	84.1	1.42
QUADRUPLE AP2α	0.18	14.74	83.67	1.41

4. Discussion

In this work we used *in silico* conformational search to analyze intrinsically disordered regions (IDRs) of two proteins participating in Clathrin-mediated endocytosis (CME) – Epsin, and Eps15. We generated *de novo* ensembles of 3-4 million feasible conformations for IDRs of Epsin and Eps15. Since the IDRs of both proteins contain multiple binding sites for the adaptor protein AP2, we wanted to investigate if such binding would alter characteristics of either of the two IDRs by restricting the available conformational space. To answer this question, we docked AP2 to the 8 binding sites in each Epsin structure, and to the 15 sites in each Eps15 structure, to obtain filtered ensembles of each IDR that would allow binding at single binding site or combinations of binding sites. We found that with an increase in the number of bound AP2 molecules, the ensemble of Epsin IDRs showed an increase in both end-to-end distance (length) and radius of gyration, (volume), which was not seen with the Eps15 ensemble.

At first glance, such contrasting results are surprising given that the IDRs of both proteins are long regions (>300 aa), have multiple binding sites for the same protein (AP2), and are

evolutionarily conserved. However, it is possible that this outcome could result from the distribution of the binding motifs along the IDR sequence. The 8 DPW motifs in the Epsin IDR sequence are moderately spaced out, with a minimum separation of 7aa and a median separation of 10 intervening residues, suggesting that AP2 binding to adjacent sites might only be possible with relatively extended conformers. In contrast, the 15 sites in Eps15 are clustered closely with a minimum of 2aa and a median of 7.5 intervening residues (with ≤ 5 intervening residues in 7 out of 14 cases). This suggests that Eps15 conformers, unlike Epsin, simply do not bind AP2 simultaneously at adjacent sites due to crowding. In other words, Eps15 conformers that bind multiple copies of AP2 would utilize non-adjacent binding sites, resulting in fewer conformational constraints imposed by AP2-binding. Having a total of 15 binding sites for AP2 might provide sufficient flexibility for accommodating AP2-binding without disallowing compact low-energy states in the Eps15 IDR. In addition, Table 1 and Table 3 show that the binding of each AP2 domain causes a ~4-10x fold decrease in the ensemble size of Epsin IDR, whereas it causes only a 1.8-2.5x fold change in the ensemble size of Eps15 IDR. These statistics corroborate the interpretation that although both IDRs bind AP2, the binding has a different impact on the IDR ensembles and on the ability of the IDRs to exert steric pressure. Our interpretation of the results is that Epsin IDR undergoes binding-induced-expansion. Such a phenomenon could allow AP2 binding to boost Epsin-derived steric pressure toward membrane bending in CME. This would build upon previous studies that found the disordered region of Epsin creates steric pressure at Clathrin-coated pits[21]. Other studies using conformational sampling of disordered regions[33,34] have also suggested a link between IDR flexibility and physical pressure, force generation, or mechanical sensing. Future experiments using techniques like SAXS on the Epsin-IDR and Eps15-IDR regions might serve to validate the hypothesis of binding-induced expansion. In addition, experiments such as Surface Plasmon Resonance or Isothermal calorimetry can shed light on the thermodynamics of AP2 binding to the IDRs.

In our study, we used large ensembles of *de* novo structures generated through random sampling of torsional angles to generate sterically-feasible structures. However, it is important to note that large ensemble sizes and conformational diversity come at the expense of structure refinement using minimization or advanced MD techniques. MD with appropriate force fields could be used to simulate structural constraints imposed by inter-residue contacts, generating a non-redundant set of biologically-selected structures[35]. However such advanced techniques are computationally prohibitive for large IDRs with high conformational entropy (both Epsin and Eps15 IDRs are at the C-terminus of their sequences, lending greater flexibility). Hence, when we explore the impact of binding on the IDR ensembles, our study uses metrics of size and energy that are appropriate for *unrefined and unminimized* protein structures. Such *in silico* studies may be considered crude estimates of reality, but they can guide subsequent experiments by prioritizing which proteins to study. When ensemble modeling is guided by experimental measurements such as residual dipolar couplings from NMR, or diffraction data from SAXS, then the *in silico* generated ensembles can realistically utilize a smaller number of structures with energy-minimization, and the analysis of the ensemble might infer subtle biochemical insights. Without experimental input, *de novo* ensembles generated from mere sequence information can still provide a first order approximation of the most prominent features of the accessible conformation space. In the case of Epsin and Eps15, even a

crude estimate of the bound and unbound ensembles showed an important distinction in how binding affects the steric volume. Binding multiple copies of AP2 appears to cause expansion of Epsin. The AP2-binding function of Eps15 is not understood, but our Eps15 results are consistent with a previous hypothesis that the IDRs of CME could serve to recruit other CME proteins from the cytoplasm[9].

In conclusion, we have used *de novo* ensembles of disordered regions to explore the novel hypothesis of binding-induced-expansion. Our theoretical results with AP2-binding to Epsin and Eps15 indicate that Epsin is a good candidate (and Eps15 is not) for future experimental testing of binding-induced-expansion. Future experiments should also test whether the spacing between binding sites modulates the impact of binding on steric volume, and steric pressure in regulating CME. The fundamental biophysics of binding-induced-expansion and ensemble-elongation may be applicable to multiple proteins, pathways, and morphogenic processes. Our interpretation of the computational findings is that binding-induced-expansion is an additional biological function performed by disordered polypeptides, and a viable mechanism for entropic forces to be leveraged and regulated.

Acknowledgements

This work is supported by the Singapore Ministry of Health's National Medical Research Council (NMRC) under its Open Fund Individual Research Grant scheme (OFIRG15nov062), and by startup funds from a Duke-NUS SRP Phase 2 Research Block Grant.

References

1. Oldfield, C. J. & Dunker, A. K. Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. *Annu. Rev. Biochem.* **83**, 553–584 (2014).
2. Dyson, H. J. Making Sense of Intrinsically Disordered Proteins. *Biophys. J.* **110**, 1013–1016 (2016).
3. Varadi, M. & Tompa, P. The Protein Ensemble Database. *Adv. Exp. Med. Biol.* **870**, 335–349 (2015).
4. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2015).
5. Berlow, R. B., Dyson, H. J. & Wright, P. E. Expanding the Paradigm: Intrinsically Disordered Proteins and Allosteric Regulation. *J. Mol. Biol.* **430**, 2309–2320 (2018).
6. Simon, J. R., Carroll, N. J., Rubinstein, M., Chilkoti, A. & López, G. P. Programming molecular self-assembly of intrinsically disordered proteins containing sequences of low complexity. *Nat. Chem.* **9**, 509–515 (2017).
7. Babu, M. M., van der Lee, R., de Groot, N. S. & Gsponer, J. Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol.* **21**, 432–440 (2011).
8. Uversky, V. N. Intrinsically disordered proteins in overcrowded milieu: Membrane-less organelles, phase separation, and intrinsic disorder. *Curr. Opin. Struct. Biol.* **44**, 18–30 (2017).
9. Dafforn, T. R. & Smith, C. J. I. Natively unfolded domains in endocytosis: hooks, lines and linkers. *EMBO Rep.* **5**, 1046–1052 (2004).
10. Babu, M. M., Kriwacki, R. W. & Pappu, R. V. Versatility from Protein Disorder. *Science* **337**, 1460–1461 (2012).
11. Eliezer, D. Biophysical characterization of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **19**, 23–30 (2009).
12. Liu, Y., Wang, X. & Liu, B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief. Bioinform.* **20**, 330–346 (2019).

13. Feldman, H. J. & Hogue, C. W. A fast method to sample real protein conformational space. *Proteins* **39**, 112–131 (2000).
14. Feldman, H. J. & Hogue, C. W. V. Probabilistic sampling of protein conformations: new hope for brute force? *Proteins* **46**, 8–23 (2002).
15. Ozenne, V. *et al.* Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinforma. Oxf. Engl.* **28**, 1463–1470 (2012).
16. Mittag, T. & Forman-Kay, J. D. Atomic-level characterization of disordered protein ensembles. *Curr. Opin. Struct. Biol.* **17**, 3–14 (2007).
17. Krzeminski, M., Marsh, J. A., Neale, C., Choy, W.-Y. & Forman-Kay, J. D. Characterization of disordered proteins with ENSEMBLE. *Bioinformatics* **29**, 398–399 (2013).
18. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
19. Huang, J. & MacKerell, A. D. Force field development and simulations of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **48**, 40–48 (2018).
20. Keul, N. D. *et al.* The entropic force generated by intrinsically disordered segments tunes protein function. *Nature* **563**, 584–588 (2018).
21. Busch, D. J. *et al.* Intrinsically disordered proteins drive membrane curvature. *Nat. Commun.* **6**, 7875 (2015).
22. McMahon, H. T. & Boucrot, E. Molecular mechanism and physiological functions of clathrin-mediated endocytosis. *Nat. Rev. Mol. Cell Biol.* **12**, 517–533 (2011).
23. Ford, M. G. J. *et al.* Curvature of clathrin-coated pits driven by epsin. *Nature* **419**, 361–366 (2002).
24. Jakobsson, J. *et al.* Role of epsin 1 in synaptic vesicle endocytosis. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 6445–6450 (2008).
25. Tebar, F., Sorkina, T., Sorkin, A., Ericsson, M. & Kirchhausen, T. Eps15 is a component of clathrin-coated pits and vesicles and is located at the rim of coated pits. *J. Biol. Chem.* **271**, 28727–28730 (1996).
26. Carbone, R. *et al.* eps15 and eps15R are essential components of the endocytic pathway. *Cancer Res.* **57**, 5498–5504 (1997).
27. van Bergen En Henegouwen, P. M. Eps15: a multifunctional adaptor protein regulating intracellular trafficking. *Cell Commun. Signal. CCS* **7**, 24 (2009).
28. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382–W388 (2005).
29. Kalthoff, C., Alves, J., Urbanke, C., Knorr, R. & Ungewickell, E. J. Unusual structural organization of the endocytic proteins AP180 and epsin 1. *J. Biol. Chem.* **277**, 8209–8216 (2002).
30. Cupers, P., ter Haar, E., Boll, W. & Kirchhausen, T. Parallel dimers and anti-parallel tetramers formed by epidermal growth factor receptor pathway substrate clone 15. *J. Biol. Chem.* **272**, 33430–33434 (1997).
31. Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43**, W389–W394 (2015).
32. Chandramohan, A. *Structural insights into folded, unfolded and nascent protein states using ensemble sampling and cluster expansion*. (2014). URL: https://scholarbank.nus.edu.sg/handle/10635/77759
33. Zhao, C., Liu, C., Hogue, C. W. V. & Low, B. C. A cooperative jack model of random coil-to-elongation transition of the FH1 domain by profilin binding explains formin motor behavior in actin polymerization. *FEBS Lett.* **588**, 2288–2293 (2014).
34. Liu, C., Yao, M. & Hogue, C. W. V. Near-membrane ensemble elongation in the proline-rich LRP6 intracellular domain may explain the mysterious initiation of the Wnt signaling pathway. *BMC Bioinformatics* **12 Suppl 13**, S13 (2011).
35. Marchetti, J., Monzon, A. M., Tosatto, S. C. E., Parisi, G. & Fornasari, M. S. Ensembles from Ordered and Disordered Proteins Reveal Similar Structural Constraints during Evolution. *J. Mol. Biol.* **431**, 1298–1307 (2019).