LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-230082

Intra-molecular cohesion of coils mediated by phenylalanine-glycine motifs in the natively unfolded domain of a nucleoporin

V. V. Krishnan, E. Y. Lau, J. Yamada, D. P. Denning, S. S. Patel, M. E. Colvin, M. F. Rexach

April 19, 2007

PLoS Computational Biology

Intra-molecular cohesion of coils mediated by phenylalanine-glycine motifs in the natively unfolded domain of a nucleoporin

V.V. Krishnan[1,2,6], Edmond Y. Lau[3], Justin Yamada[2], Daniel P. Denning[4], Samir S. Patel[2], Michael E. Colvin[5] and Michael F. Rexach*[2]

[1]Department of Applied Science
University of California, Davis
Davis, California

[2]Department of Molecular, Cell and Developmental Biology
University of California, Santa Cruz
Santa Cruz, California

[3]Biology and Biotechnology Division
Lawrence Livermore National Laboratory
Livermore, California

[4]Department of Biology
Massachusetts Institute of Technology
Cambridge, Massachusetts

[5]Center for Computational Biology
School of Natural Sciences
University of California, Merced
Merced, California

[6]Department of Chemistry
California State University, Fresno
Fresno, CA 93740

*Correspondence should be addressed to M.R. *email: rexach@biology.ucsc.edu*

Runing Title: Dynamic ensemble of structures of a nucleoporin FG domain

Abbreviations: NPC, nuclear pore complex; FG nup, nucleoporin with FG motifs; FG domain, nucleoporin domain with FG motifs; MD, molecular dynamics; NMR, nuclear magnetic resonance

Abstract:

The nuclear pore complex (NPC) provides the sole aqueous conduit for macromolecular exchange between the nucleus and cytoplasm of cells. Its conduit contains a size-selective gate and is populated by a family of NPC proteins that feature long natively-unfolded domains with phenylalanine-glycine repeats. These FG nucleoporins play key roles in establishing the NPC permeability barrier, but little is known about their dynamic structure. Here we used molecular modeling and biophysical techniques to characterize the dynamic ensemble of structures of a representative FG domain from the yeast nucleoporin Nup116. The results show that its FG motifs function as *intra*-molecular cohesion elements that impart order to the FG domain. The cohesion of coils mediated by FG motifs in the natively unfolded domain of Nup116 supports a type of tertiary structure, a native pre-molten globule, that could become quaternary at the NPC through recruitment of neighboring FG nucleoporins, forming one cohesive meshwork of intertwined filaments capable of gating protein diffusion across the NPC by size exclusion.

Introduction:

The nuclear pore complex is a supramolecular protein structure in the nuclear envelope that controls nucleo-cytoplasmic traffic and communication. Its most important architectural feature is a poorly understood semi-permeable diffusion barrier at its center, which allows passive diffusion of particles less than 3-4 nanometers in diameter and opens to allow facilitated transport of larger particles up to 40 nm in diameter. The NPC is composed of thirty proteins or nucleoporins (Nups) that are present in multiple copies. Among these, a group that contains numerous phenylalanine-glycine repeats (FG nups) populates the transport conduit of the NPC (**Fig. 1A**). These FG nups are thought to function as stepping-stones for karyopherin movement across the NPC [1, 2] and as structural elements of the NPC permeability barrier [3].

The three dimensional structure of *S. cerevisiae* FG nups is unusual because their large 150-700 amino acid (AA) FG domains are unfolded in their native functional state [4]. Since there are ~175 FG nups in the transport conduit of each NPC, it is currently hypothesized that the conduit is occupied and/or flanked by 175 filamentous, natively unfolded FG domains, which could 'reach-out' 50 to 350 nanometers, if fully extended. Together these domains constitute more than 17% of the total NPC mass, or >8.5 MDa of its ~50 MDa structure in yeast. Whether these FG nup filaments function as repulsive bristles that create an entropic gate at the NPC entrance [5, 6], or as cohesive filaments that form a selective meshwork at the NPC [7], has been a subject of much speculation. Recently however it was shown that the FG domains of most FG nups anchored at the NPC center interact with each other weakly via hydrophobic attractions between FG motifs to form a cohesive meshwork of filaments, whereas most FG domains of nups anchored at the periphery of the NPC do not form such cohesions and may instead behave more as entropic bristles [3]. Although there are three subtypes of FG nup filaments defined by their content of FxFG, GLFG or SAFGxPSFG motifs, their ability to interact with each other (*i.e.* their cohesiveness) appears to depend on the AA composition between FG motifs rather than on the specific FG motif [3].

It is generally assumed that natively unfolded proteins, including the FG domains of nups, have *some* preferred 3-D structure(s) dictated by intra-molecular cohesion [8, 9]. Current evidence that the FG domains of nups have *some* structures is based on CD and FTIR spectroscopic analysis, which indicates that FG domains have anywhere from 5 to 20% alpha helical and beta-sheet content at any given moment [4], though the location of such structures in the protein is probably ever changing. Indeed, the conformational flexibility inherent to natively unfolded proteins and protein domains such as those in the FG nups, places them beyond the

reach of classical structural biology tools such as X-ray crystallography and homology-based computational methods [10-13]. However, it is clear that these and other unfolded proteins participate in a wide range of key cell biological processes [14, 15] and that their native plasticity bestows specific functional properties, such as rapid molecular interaction times and the ability to bind multiple proteins simultaneously [16]. In the case of nucleoporin FG domains, a key function is to bind multiple karyopherins [17] with very rapid interaction times [18]. Thus, in contrast to folded proteins, the structure of natively-unfolded proteins must be described by an ensemble of inter-converting conformers.

Since traditional experimental methods for elucidating protein structure cannot be used with natively-unfolded proteins, new approaches are needed to study and describe their dynamic ensemble of structures. In this emerging area of research, Jha et al [19, 20] have recently introduced a general statistical coil model and Bernado et al [21, 22] have estimated the NMR measured residual dipolar couplings (RDCs) [23] from dynamic simulations to characterize the ensemble-averaged conformations of α-synuclein. Also, Ollerenshaw et al [24] have applied a native-centric topological model to understand the essential folding/unfolding dynamics SH3 domains, and Pappu and co-workers have characterized poly-glutamines as a function of chain-length conformational sampling by molecular dynamics (MD) and Monte Carlo simulations [25]. Most of these computational investigations suggest the existence of a preferred ensemble of conformers for each protein, rather than suggesting pure random coils.

Here we conducted multiple molecular dynamics simulations and biophysical measurements on a representative FG domain from the yeast nucleoporin Nup116 to test the notion that phenylalanines in its FG motifs function as structural *intra*-molecular cohesion elements. Apart from its cell biological significance, we chose this protein as a model system to investigate how a combination of molecular dynamics simulations and clustering methods can be used to characterize the ensemble of structures adopted by natively unfolded proteins, such as the FG domain of nups.

Results:

In the analysis that follows we first used MD simulations to generate a statistical ensemble of coil conformations for a 110 AA region of the Nup116 FG domain containing ten FG motifs (wild type), and of a mutant version thereof lacking the phenylalanines in the ten FG motifs (F>A mutant) (**Fig. 1B**). The MD trajectories were then analyzed to evaluate the degree of secondary structure, the overall dimensions of the protein conformations, and the contribution of the FG motifs to the intra-molecular cohesion of coils (i.e. compaction) in the dynamic ensemble of nup structures. The simulated FG domains were then expressed in bacteria, purified to homogeneity, and analyzed by NMR spectroscopy and sieving columns to quantify their overall shape through measurements of diffusion coefficient and Stokes radii. Finally, the computational and biophysical analyses were compared and combined to estimate the tertiary structure that best describes the GLFG-rich domain of a representative FG nucleoporin.

Molecular Dynamics Simulations:

Twenty independent MD simulations were performed at 300°K (25°C) on the wild type (6 ns) and F>A mutant (5 ns) versions of the Nup116 FG domain starting from a fully-extended conformation. The goal of these simulations was to sample the conformational distribution of the proteins as close as possible to their native distribution in solution. As soon as the simulations started, within the first 100 picoseconds (ps), the extended FG domains collapsed into a more cohesive or compact ensemble of structures with small patches of unstable (see below) secondary structure (**Fig. 1A** has a subset of 'end'-structures). Since the wild type and mutant FG domains are highly flexible and disordered, the resulting end structures from each of the simulations did not resemble one another, as was expected for natively-unfolded proteins. Despite the fact that the nup structures were 'ever-changing' (see below), the ensemble of structures for each did reach a dynamic equilibrium early in the simulation, because the degree of structural variability during the last 3 nanoseconds (ns) did not change significantly. This was evidenced by their constant radius of gyration (**Fig. S7**) and its statistical analysis using Tukey's Honest Significant Difference method (data not shown), which showed no significant change in the range of Rg values during the last 3 ns.

To describe quantitatively the structural dynamics of the FG domains, we calculated the auto-correlation function of a vector of the 118 phi and 118 psi angles along the peptide backbone of FG domain structures sampled every 1 picosecond (ps) from the MD trajectory. The auto-correlation functions with a 200 ps window from the final 3 ns of simulation of all

twenty wild type and F>A mutant FG domain simulations are shown, along with the comparable auto-correlation function from the MD simulation of a control protein that is folded (fibroblast growth factor 1) (**Fig. S8**). We found that for each of the replicate nup simulations the correlation in the phi-psi angles dropped from 1 to 0.738 (± 0.031) or 0.741 (± 0.023) in 1 ps for the wild-type or mutant FG domains, respectively, and then slowly decayed over 200 ps to 0.672 (± 0.037) or 0.665 (± 0.029). In contrast, for the fibroblast growth factor 1, the auto-correlation function dropped to only 0.875 in 1 ps, and then to 0.864 over the 200 ps auto-correlation window. These results indicate that the wild type and mutant FG domains are highly dynamic in comparison to folded proteins, and are constantly changing their structure.

Structural Analysis of Molecular Dynamics Simulations:

The ensemble of structures for each of the twenty MD trajectories generated for the wild type and F>A mutant FG domains were sampled from the final 3 ns of the MD simulations, at 1 ps intervals, to yield a total of 60,000 structures for each protein. The secondary structure content was then analyzed to determine the fraction of time during the simulations that each residue spent as part of a "helical" structure (either an alpha-helix or a 3_{10}-helix). In general, we found no significant difference in overall helical content between wild type and mutant FG domains. However, the alpha- and 3_{10}- helical structures that did form, ranged in size from 2-6 AA residues and did not persist for more than 15 and 35 ps on average, respectively (data not shown). The maximum duration of an alpha helix and a 3_{10}-helix was 97 and 699 ps, respectively (data not shown).

Using the same set of 60,000 structures we calculated two measures of protein compactness, the end-to-end distance between the terminal residues and the radius of gyration. The average (± 1 standard deviation) end-to-end distance for the wild type FG domain was 20.42 Å (± 9.51) compared to 20.69 Å (± 7.78) for the mutant (data not shown), and the predicted radius of gyration (R_g) was 14.52 Å (± 1.18) for the wild type, and 14.41 Å (± 1.24) for the mutant FG domain (**Fig. 2**). The different simulations sampled different regions of conformation space because significant run-to-run variations were observed in the probability distributions for each structural parameter. The similar R_g and end-to-end distance values obtained for the wild type and mutant FG domains implied that both proteins occupy equivalent hydrodynamic volumes. However previous researchers had reported that increasing the temperature to 340°K of MD simulations could improve agreement with NMR protein

conformational studies [26]. We therefore extended our MD simulations for an additional 1 ns at 325°K (52°C) or at 350°K (77°C), and a very different picture emerged (**Fig. 2**).

First, at 325°K there was a slightly greater difference in the average radius of gyration between the wild type (15.11 ± 1.43 Å) and the mutant (15.76 ± 2.58 Å) FG domains. Five of the twenty mutant simulations now had an average R_g greater than 18 Å, but all the wild type simulations had an average R_g below 18 Å. Also, the average end-to-end distance for the wild type FG domain was 20.84 Å (±10.75) compared to 24.26 Å (±13.16) for the mutant. The greater end-to-end distance for the mutant domain was likely due to thermal denaturation during the simulation.

Second, at 350°K there was a greater difference in the radius of gyration (R_g) for the wild type and mutant Nup116 FG domains (**Fig. 2** and **Fig.3**). The average R_g was 17.40 Å (± 3.11) for the wild type and 23.68 Å (± 6.05) for the mutant. Again, the F>A mutant FG domain was more sensitive to heat denaturation than the wild type. At 350°K, fifteen of the twenty mutant simulations had an average R_g greater than 20 Å, compared to only three for the wild type simulations. Consistently, the average end-to-end distance for the wild type FG domain was 29.95 Å (± 16.04) compared to 52.56 Å (± 25.31) for the mutant. These data combined provide a first indication that the F>A mutant Nup116 FG domain is not as intra-molecularly cohesive and compact as the wild type version.

Intra-molecular distances between FG motifs in the Nup116 FG domain:

As a way of assessing the dynamic structure of the FG domain, particularly from the point of view of the FG motifs, we plotted the distances between the backbone β-carbons (Cβ) for the ten sites that correspond to the phenylalanine or to the substitute alanine residues in the various FG motifs. The distances used were from the MD simulations at 350°K, which most closely resembled the physical measurements obtained later (see below). The distance analysis yielded 45 F-F (or A-A) distances for each structure (**Table S2**). Probability distributions for each Cβ to Cβ distance were calculated and analyzed looking for significant differences between the wild type and mutant FG domains. To estimate the sharpness of the Cβ to Cβ distance distributions we counted the number of 1 Å wide bins that had greater than 10% of the probability distribution (no bin had more that 20% of the probability distribution). This metric was calculated for all 45 Phe-Phe Cβ to Cβ distances in all twenty replicates of the FG domain simulations. In stable tertiary structures, these distances occur as sharp peaked distributions around the equilibrium inter-residue distance; in a fully random ensemble of structures, they

occur as a very broad distribution; and in semi-structured proteins, they occur as one or more intermediate-width distributions. For the wild type FG domain simulations, nine of the ten simulations analyzed had more than 3 sharp peaks. By comparison, only three of the ten F>A mutant simulations exhibited a similar level of sharpness in the distance distributions (data not shown). This provides tentative evidence that the wild type FG domain is more structured than the F>A mutant. Repeating this analysis to include only peaks with the distance distributions of <15 Å yielded a very similar result (data not shown). Representative examples of the probability distribution obtained from the MD trajectories at 350°K are shown in **Figure 4A.** The values correspond to the distribution of distances between F84 and F93 (see **Fig. 1B**) in the wild type FG domain, and between A84 and A93 in the F>A mutant domain. To allow comparison of the average inter-residue distances, the probability distributions were fit to a single Gaussian distribution, though in some cases there are multiple distinct peaks. This is only a rough approximation to the observed probability distribution, but the assumption can be justified in the context that these ensemble of structures are to be used (see below), after all, these structures are rapidly inter-converting and the width of the Gaussian is broad enough to accommodate all of the major peaks in the distribution. For example, in the case of the F84-F93 distance distribution a Gaussian centered at 16.2 Å with a width of 11 Å covers both peaks at 10 and 20 Å (**Fig. 4A**).

Probability distributions obtained from the MD simulations were subjected to clustering using the Pearson squared correlation. This was done to determine how any two of the distributions sampled in regular intervals of the MD simulations are correlated with each other. The correlation coefficient does not depend on the specific measurement units used because other correlation coefficients such as Euclidian distance metric yielded similar clustering effects (data not shown). **Figure 4B** shows the matrix intensity plots of the correlations for the wild type and mutant FG domains. The indices of the matrix correspond to the various Phe-Phe pairs (or Ala-Ala pairs) as listed in **Table S2**. Indices 1 through 9 correspond to the distance from the first Phe (at position 13) to the other 9 Phe (positions 23 through 113), while indices 10-18 correspond to similar distances from the second Phe (at position 23) to the other eight Phe's (positions 32 through 113) and so on. Altogether, the clustering analysis showed that there is a stronger correlation between the various Phe-Phe distributions in the wild type FG domain than between the various Ala-Ala distributions in the F>A mutant. This indicates that the wild type FG domain is more ordered than the mutant.

To better describe the relationship between FG motifs in the FG domain, the distances between F-F pairs (or substitute A-A pairs) were categorized into groups representing distances of 10-15 Å, 15-20 Å or > 20 Å, which are color coded in **Figure 4C** as thick, medium or thin lines, respectively. A list of all 45 distances for both proteins is given in **Table S3**. Among the F-F distances in the wild type FG domain, seven F-F pairs are less than 15 Å apart, (**Table S3** and **Fig. 4C)**. In contrast, the F>A mutant FG domain had only four A-A pairs with such short distances. In the wild type FG domain, four F-F pairs were in the range of 15-20 Å apart, while seven F-F pairs were farther than 20 Å. In the F>A mutant, there were five A-A pairs with distances in the mid-range (15-20 Å), and four pairs showing distances greater than 20 Å. This analysis demonstrates that the *intra*-molecular distances between FG motifs in the wild type and mutant FG domains are quantifiably different from each other. From the analysis, we conclude that there is a tendency for FG motifs to be proximal in the wild type domain, which is absent in the mutant. This observation is consistent with the notion that FG motifs interact intra-molecularly.

Self-diffusion coefficient measurements of the purified Nup116 FG domain by NMR:

The structural predictions made by the *in silico* modeling prompted us to seek physical evidence that the phenylalanine residues in FG motifs function as structural cohesion elements through *intra*-molecular interactions within the Nup116 FG domain. In principle, a change in the dynamic ensemble of structures of the FG domain (resulting from the substitution of all Phe's for Ala) could be detected by NMR, as a less-ordered mutant FG domain would exhibit a slower diffusion coefficient. Wild type and mutant versions of the Nup116 FG domain were purified to homogeneity and subjected to NMR analysis.

Figure 5A shows the plot of the one-dimensional ^{1}H NMR spectra for the wild type and mutant Nup116 FG domains at 600 MHz, at a probe temperature of 25°C. Due to the lack of folded structures in the FG domains, it was anticipated that their hydration would be significantly different from that of ordered, globular proteins [29, 30]. When pre-saturation of the water was used, we found significant reduction in the intensity of the amide region of the spectrum due to fast exchange with the solvent protons [27, 28]. This observation, combined with the narrow chemical shift dispersion of the amide resonances (7.9-8.5 ppm) in both spectra, is a clear indication that both FG domains are natively unfolded and highly dynamic. NMR experiments were also performed at lower temperatures (5° and 10°C), but no significant improvements in the spectral dispersion were noted (data not shown).

Figure 5B shows the experimental self-diffusion measurements (intensity vs. product of the area of gradient pulse strength and the diffusion length) of the FG domains and the corresponding exponential fits. The data yielded self-diffusion coefficients (D_s^{expt}) values of 13.17 ± 0.26 and 12.18 ± 0.12 x $10^{-11} m^2 s^{-1}$ for the wild type and mutant FG domains, respectively, indicating slower diffusion for the less ordered, mutant FG domain. As expected, the diffusion of both FG domains (which are ~12.5 kDa in size in their purified form) is significantly slower than a folded protein of higher molecular weight such as lysozyme (MW = 14.1 kDa; D_s = 10.9 x10^{-11} $m^2 s^{-1}$) [31], indicating that the wild type and mutant FG domains have extended structures that sample a relatively large conformational space, as other unfolded proteins do [32]. The mass and volume of the F>A mutant domain is smaller than wild type due to the replacement of 10 Phe to Ala, yet the diffusion constant of the mutant was smaller on average, suggesting that its effective hydrodynamic volume is larger.

Characterization of the Nup116 FG domain in sieving columns:

To further characterize the hydrodynamic properties of the wild type and mutant Nup116 FG domains we analyzed them in sieving columns to determine their Stokes radii. We expected that a less ordered mutant FG domain would occupy more hydrodynamic space and would elute faster from the column. Purified wild type and mutant versions of the Nup116 FG domain were subjected to size-fractionation through an FPLC Superdex 75 column, and their elution profile was compared to those of carbonic anhydrase (29 kDa, Rs = 23.5Å), ovalbumin (45 kDa, Rs = 29.8Å) and BSA (68 kDa, Rs = 35.6Å), which served as size standards. The observed Stokes radius for the wild type FG domain was 25.6 Å (**Fig. 6A**); this is larger than carbonic anhydrase, despite the FG domain (12.5 kDa) being less than one-half the size. This highlights the fact that the FG domain is natively unfolded. The F>A mutant FG domain eluted faster from the sieving column than its wild type counterpart and migrated with a Stokes radius equivalent to 27.6 Å (**Fig. 6A**). This apparent loss of compaction (*i.e.* the hydrodynamic radius of the mutant FG domain being larger that the wild type FG domain) was expected based on the slower NMR diffusion coefficient for the mutant, and based on the computationally predicted difference in hydrodynamic volumes of the wild type and mutant FG domains at 350°K. This loss of intra-molecular cohesion supports the notion that phenylalanines in the natively unfolded FG domain interact intra-molecularly (and may cluster) through hydrophobic attractions. Indeed, when the sizing column was pre-equilibrated in buffer with a high salt concentration (1M KOAc), the wild type FG domain became more compact than it was at physiological salt (150 mM KOAc) and

eluted later from the column with a Stokes radius equivalent to carbonic anhydrase, which itself did not change compaction in high salt (**Fig. 6B**). The F>A mutant also compacted a little in the high salt, suggesting some residual hydrophobic interactions (perhaps through leucine residues in the GLAG motifs), but the compaction was much more subtle than observed for the wild type FG domain. We conclude that intra-molecular hydrophobic interactions between phenylalanine residues in GLFG motifs bias the arrangement of coils within the FG domain to form an ensemble of dynamic, non-random tertiary structures with a quantifiable level of *intra*-molecular cohesion. Finally, the Stokes radii measured for the wild type and mutant FG domains matched well with their predicted radius of hydration obtained by computational modeling at 350°K (**Table 1**).

The measured hydrodynamic volume of the Nup116 FG domain predicts a native pre-molten globular structure:

The hydrodynamic volume of a protein in different structural configurations (*i.e.* a folded globule, a molten globule, a pre-molten globule, a random coil, etc.) can be predicted from its mass using simple equations [33], These equations were derived from the analysis of large data sets containing experimentally-determined hydrodynamic values for proteins in various structural configurations. Based on the mass predicted for the wild type and mutant FG domains, we calculated the hypothetical hydrodynamic volumes that each would occupy in various structural configurations (**Table 1**), and then compared the predicted values to our experimentally-determined values aiming to identify the predicted structural configuration for each FG domain that best matched our experimental results. For the wild type FG domain, a pre-molten globule structure matched best its measured hydrodynamic volume (**Table 1**). Indeed, a similar structure is adopted by natively-unfolded proteins with high hydrophobicity [33]. For the F>A mutant, its predicted structure had a more extended configuration, and matched best the structure that proteins adopt in 8 M urea [33]. A similar structure is also adopted by natively unfolded proteins with low hydrophobicity [33]. From this analysis we concluded that phenylalanine residues in GLFG-rich domains function as intra-molecular cohesion elements capable of compacting the natively unfolded structure of an FG domain from an extended coil-like configuration to a more compact, pre-molten globule like configuration.

Discussion:

We have demonstrated an effective computational approach to characterize the dynamic ensemble of structures adopted by a natively unfolded protein. We characterized the ensemble of conformations adopted by the natively unfolded FG domain of the *S. cerevisiae* nucleoporin Nup116 (AA 348-458), and of a mutant version thereof lacking the phenylalanines in its predominantly GLFG motifs (**Fig. 1**). Both FG domains were found to be highly dynamic and disordered yet contained quantifiable physical differences between them. Specifically, the MD simulations at 350°K (**Fig. 2**) predicted a more cohesive and/or compact ensemble of structures for the wild type FG domain compared to the F>A mutant based on the average radius of gyration and end-to-end distances (**Fig. 3**). This structural prediction was supported by the inter-phenylalanine (or inter-alanine) distance analysis (*i.e*. the distance between wild type or mutant FG motifs, respectively) and the Pearson correlations of F-F (or A-A distances) in the FG domains (**Fig. 4**), which indicate that the FG domain has increased probability of sampling geometries that are more ordered when the phenylalanines in FG motifs are present. Importantly, the general structural predictions made by these computational analyses were confirmed by direct physical examination of wild type and mutant FG domains by NMR based measurement of their hydrodynamic properties (**Fig. 5**) and by their hydrodynamic radius in a sieving column (**Fig. 6**).

The measured Stokes radii for the wild type (Rs = 25.6 Å) and mutant (Rs = 27.6 Å) FG domains matched best the hydrodynamic volumes predicted by the MD simulations when these were preformed at 350°K (**Table 1**) rather than when performed at 300°K, which is more common. The values predicted at room temperature (300°K) were significantly smaller than the dimensions of the FG domains measured in the sizing column. Also, the 300°K simulations did not predict the significant difference in hydrodynamic volume between the mutant and wild type FG domains, which we detected by NMR and sizing column measurements. We suggest that MD simulations of natively-unfolded proteins at higher temperatures (350°K) are a more accurate predictor of the physical dimensions of natively-unfolded proteins in solution.

The evidence gathered here suggests that GLFG motifs function as cohesive structural elements that bias the arrangement of coils within the FG domain of Nup116 and condense its filamentous, dynamic structure into more cohesive, less disordered states. Presumably, the GLFG motifs could exert their structural influence through hydrophobic pairing, stacking, zippering, or otherwise clustering of the aromatic ring of phenylalanine side chains through

energetically favorable aromatic edge-to-face interactions, as opposed to less favorable face-to-face (π–π) interactions [34]. Moreover, the leucine residue in the GLFG motif may enhance the Phe clustering effect by increasing the local hydrophobicity or by influencing the orientation of the neighbor Phe ring. A two dimensional representation of the FG domain AA sequences in a hydrophobic cluster analysis (HCA) [35, 36] illustrates the hydrophobic patches in the FG domain surface best (**Fig. S9**). Although HCA is most useful in determining hydrophobic clusters in helical patterns (*e.g.* helical structure for transmembrane segments) [31] it is still informative in the absence of a structural fold, because it allows the identification of hydrophobic patches between nearby AAs in a protein sequence. Indeed, the HCA analysis clearly indicated that LF pairs in GLFG motifs form hydrophobic patches in the wild type FG domain, and unexpectedly, that methionine-Phe patches could also form. Both of these hydrophobic patches were missing in the F>A mutant domain (**Fig. S9**). Thus, the F>A mutant FG domain is less compact probably because it does not have hydrophobic patches that can make intra-molecular hydrophobic interactions. Also, the lack of hydrophobic patches may also explain why this mutant cannot form inter-molecular interactions with other GLFG domains, as the wild type FG domain does [3].

Although the observed phenylalanine-mediated stabilization of FG domain structures is uncommon, a recent report by Dhe-Paganon et al., defined a 'phenylalanine zipper' motif within the hydrophobic core of APS, which is critical for APS dimerization [37]. In that case, the aromatic side chains of ten phenylalanine residues are uniquely stacked to form a zipper that is stabilized by helical secondary structures in the APS backbone. A detailed NMR analysis of the Nup116 FG domain will be needed to establish whether the Phe's in its FG motifs form similar zipper structures. However, if such zipper structures existed within FG domains of nucleoporins, they would be unstable and highly dynamic given the lack of secondary structure in the FG domains. Other types of nucleoporin FG domains, such as those containing mostly FxFG motifs (*e.g.* in nups of the yeast nuclear basket structure) or SAFGxPSFG motifs (*e.g.* in nups of the yeast cytoplasmic fibers) may display different *intra*-molecular interactions from those observed here for the GLFG-rich domain of Nup116. Indeed, the isolated FxFG domain from a vertebrate nucleoporin (nup153) appears to lack *intra*-molecular interactions [38].

The finding that GLFG motifs function as *intra*-molecular cohesion elements that impart cohesion to the natively unfolded domain of nups has important implications for the general architecture of the NPC *in vivo*. Indeed, this structural feature (*i.e.* cohesion between FG motifs) may be prevalent among yeast nups that contain predominantly GLFG motifs, such as Nup116,

Nup100, Nup49, Nup57 and nNup145 **(Fig. 1**). These nups are anchored to the NPC center and their GLFG domains combined account for >8.4% of the total NPC mass, or >3.7 MDa of the NPC structure. There are >1,750 GLFG motifs in each yeast NPC scattered along >80 GLFG domains of nups. If these GLFG motifs were to exert their cohesive effect *inter*-molecularly (as observed [3]) as well as *intra*-molecularly (as predicted here), the >80 filamentous GLFG domains within each NPC could coalesce into a single cohesive, but highly flexible meshwork of intertwined filaments, which could gate molecular diffusion by size-exclusion (*i.e.* a filamentous sieve) [4, 7]. The cohesiveness of such meshwork would be maintained by the weak but numerous interactions between GLFG motifs [7, 39], and the sieve-size may be related to the residue spacing between the FG motifs, which is usually 10-20 AA long and is conserved among GLFG-rich FG domains in yeast. Clearly, a more detailed structural characterization of the various types of FG domains is needed to understand fully the structure and molecular dynamics of the nuclear pore complex. Notwithstanding, the available evidence indicates that there may be two types of natively unfolded FG domains operating at the NPC. Those that are *intra*- and *inter*-molecularly cohesive, which would operate collectively as a single meshwork to gate traffic through the NPC; and those that are not *intra*- or *inter*-molecularly cohesive, which would operate individually as repulsive bristles that gate molecular traffic by forming an entropic barrier to diffusion.

Based on the measured hydrodynamic volume of the Nup116 GLFG domain, the estimated number of GLFG-rich domains in the NPC, their location anchored at the NPC center, and the predicted volume of the central NPC transport conduit, we were able to estimate the predicted hydrodynamic volume occupied by GLFG-rich domains of nups at the NPC center, assuming they all exist in a pre-molten globular state. Surprisingly, these FG domains of nups (which represent 5 out of the six FG nups anchored at the NPC center) would only occupy 22,928 nm^3 of space, which is roughly one-third of the volume available to them within the transport conduit (64,821 nm^3) (**Table S4**). At face value, this implies that in a “ground state” the presumed filamentous meshwork of FG domain at the NPC center likely has a hole in the middle, assuming that the FG domains remain in a pre-molten globular shape about their tether sites. This is in fact what has been observed for isolated NPCs is some cases, where there seems to be a hole in the middle of the conduit [40]. This however assumes that the volume occupied by the centrally anchored Nsp1 (which could total up to 93,320 nm^3 according to our measurements and calculations; data not shnown), and by the other peripherally anchored FG nups (~60,000 nm^3 combined) are flanking the entry and exit sites of the transport conduit. This would seem to be necessary as the combined volume of the Nsp1 FG domains alone, or with

the other FG domains combined, would not fit in the remaining volume of the transport conduit. Finally, when we divided the remaining 'empty space' in the transport conduit by the average volume of a shuttling receptor such as importin beta, we deduced that there is enough room left over at the NPC center for ~250 importin molecules. This is very close to the number of importins that are presumed to cross the NPC simultaneously (Ref.), and is very close to the number of importin molecules that we measured can bind per NPC, using a photo-crosslinking approach between purified importin beta and isolated nuclei (S. Patel and M. Rexach, personal communication).

Acknowledgements:

This work was supported in part by NIH Grant #GM061900 awarded to M.R. This work was in part performed under the auspices of the United States Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract number W-7405-ENG-48. This work was supported in part by the U.S. Department of Energy, Office of Science, Offices of Advanced Scientific Computing Research, and Biological & Environmental Research through the U.C. Merced Center for Computational Biology.

Materials and Methods:

Classical molecular dynamics simulation: MD simulations of individual FG domains were started from a fully extended backbone structure (i.e. with the phi and psi angles set to 180 for all residues except for the three proline residues, which put a 60° bend in the sequence. A different random number seed was chosen for each of the different simulations to randomize the initial atom velocities. Twenty separate simulations were run for either 6 ns (wild type) or 5 ns (mutant) each using different initial atomic velocities and analyzed at 1 ps intervals. The wild type fragment required an additional nanosecond of dynamics to have its radius of gyration converge. All MD simulations were performed with AMBER [41-43] using implicit solvent models. Each molecule was simulated in the presence of a Generalized Born/Surface Area (GB/SA) implicit solvent model [44] that calculates an effective solvation energy as an empirical parameter multiplied by the exposed surface area of different atom types. Each molecule was simulated using the GB/SA implicit solvent implementation in Amber versions 7 and 8. Each system is energy minimized using 100 cycles of conjugate gradients. Constant temperature molecular dynamics at 300°K with a coupling constant of 0.2 ps was performed on the minimized systems using the standard partial charges for the Amber force field and Bondi radii for the atoms. Bonds containing hydrogens were constrained using SHAKE and a time step of 2 fs was used in all simulations. A cutoff of 250 Å was used for the electrostatic interactions, which for this system is equivalent to infinity. The salt concentration (Debye-Huckel screening) was set at 0.15 M. Secondary structure analysis: For the final 3 ns of each simulation, the structure was analyzed every 1 picosecond using a standard program for identifying secondary structure from atomic coordinates (Define Secondary Structure of Proteins; DSSP) [45]. Radius of gyration and end-to-end distance: For the final 3 ns of each simulation, radii of gyration and the end-to-end distances between terminal residues were calculated using the program CARNAL and ptraj, distributed with AMBER 7 [42, 43] . High temperature molecular dynamics simulations: Molecular dynamics were performed at elevated temperatures for each of the 20 wild-type and mutant simulations. The GB/SA simulations were all restarted after 5 ns coupled to a heat bath at 325°K and 350°K with all other parameters of the simulation kept the same. The simulations were run for 1 ns and the final 500 ps were used for analysis.

Autocorrelation functions: To determine the degree of dynamical change in the FG domain structure, the autocorrelation function of a vector composed of the 118 phi and 118 psi angles (total vector length = 236) along the peptide backbone of structures sampled every 1 ps from

the MD trajectory (was derived). The autocorrelation function was computed as the dot-product of successive phi-psi vectors, using every 1 ps step as a new time origin and calculating to correlation function out to 200 ps. Several different time windows were used in the autocorrelation calculation but all gave the same result. As a control, the same autocorrelation function was calculated for the first 118 phi-psi angles (out of 130) for an MD trajectory of a well-folded protein (fibroblast growth factor 1 PDB id= 1AXM).

Calculation of probability distributions: The MD trajectories of 45 F-F distances between the Cβ atoms were analyzed to calculate the probability distributions. Systematically each of the F-F distance was interrogated and the distances were binned (1 Å bins from 0 to 60 Å) to form a histogram of distance distributions. Probability distributions were calculated for the 10 simulations independently and averaged at the end. A similar procedure was adopted for the F>A mutant where the distance between the Cβ atoms of the Ala residue was used. Final probability distributions were used without any normalization.

Intra-molecular distance constraints: As a first approximation the probability distributions were fit to a Gaussian distribution (probability vs. distance). This is a conservative approach and is expected to be valid considering the number of structures generated (~100,000) during the molecular dynamics simulations and in the absence of any constraints. The center of the Gaussian is considered as the mean distance between the F-F (or A-A), while the width at half-maximum is used as the allowed variation in the constraint. Clustering analysis: The correlation between different F-F probability distribution reflects the degree to which these variables (F-F distances) are related. The most common measure of correlation is the Pearson Product Moment Correlation (www.r-project.org) and reflects the degree of linear relationship between the two variables. In order to determine whether probability profiles of the F-F interaction correlate, a similarity matrix with a Pearson square metric was calculated. The correlation was used to indicate the presence (or absence) of relationship between various F-F interactions.

Expression and purification of FG domains*:* FG domains were expressed in *E. coli* as fusion proteins with GST (glutathione S transferase) at the N-terminus and a HIS tag (six contiguous histidines) at the C-terminus. AA replacements (F>A) were generated using standard site directed mutagenesis and the changes were confirmed by DNA sequence analysis. Glutathione

coated beads were used to isolate each GST-FG domain-HIS fusion protein from crude bacterial cell extracts. The FG domain-HIS protein was then eluted from the beads by specific thrombin proteolysis of the GST tag. Nickel-coated beads were then used to capture and isolate the FG domain through its HIS tag, and the captured proteins were eluted from the beads using imidazole. Finally, the concentrated eluates were fractionated in an FPLC Superdex 200 sizing column that was equilibrated in phosphate buffer at physiological pH.

NMR experiments: NMR experiments were performed on samples dissolved in 50 mM NaH_2PO_4, pH of 6.5. Final protein concentrations were ~0.5 mM for both FG domains. NMR experiments were performed in a Varian INOVA 600 MHz spectrometer equipped with a 5mm probe with a single-axis (along Z) shielded magnetic field gradients. One dimensional ^1H NMR experiments were obtained using the water suppression scheme 1-3-3-1 Water-gate [46]. Self-diffusion coefficient measurements were obtained using a BPP-SED (bipolar-gradient pulse pair selective echo dephasing) sequence [47].

Hydrodynamic calculations: Translational diffusion tensor values were calculated based on the beads model approximation of García de la Torre and Bloomfield [48]. This method has been used successfully to calculate translational as well as rotational diffusion tensors of proteins [31, 49]. All atoms were considered as beads of equal size (σ = 5.1 Å). The overall isotropic translational self-diffusion coefficient was calculated by taking the average of the principal values of the diffusion tensor.

Calculating hydrodynamics radius from the simulated radius of gyration: The hydrodynamics radius (R_h) for the WT and mutant Nup116 fragment was calculated from the radius of gyration (R_g) from the simulations. This can be accomplished by using the scaling relationship give in {Wilkins, 1999 #5177}. For native proteins, the scaling relationship is $R_g = 0.77R_h$ and for protein in strongly denaturing conditions the scaling relationship is $R_g = 1.06R_h$. For the wild type and mutant Nup116 FG domains simulated at 300° and 325°K, the hydrodynamics radius was obtain by $R_h = R_g/0.77$. In the 350°K simulations, some of the protein conformations are highly extended (as in denaturing conditions) and a single scaling value is not appropriate. In this case, if the R_g for a structure was less than 27.3 Å for wild type and 26.3 Å for the mutant, it was scaled by 1/0.77 and if R_g were greater the values it was scaled by 1/1.06. The R_g cutoff values

of 27.3 Å (WT) and 26.3 Å (mutant) were obtained by using Uversky's relationship: R_h (8 M urea) = (0.22)*$M^{0.52}$, where M is the molecular mass [33]. The molecular mass for the wild type FG domain was 11,900 daltons and the mutant fragment was 11,100 Daltons. R_g was obtained by scaling R_h by 1.06, *i.e.* R_g (WT, urea) = [0.22*$(11900)^{0.52}$]/1.06.

References:

1. Rexach, M. and G. Blobel, *Protein import into nuclei: association and dissociation reactions involving transport substrate, transport factors, and nucleoporins.* Cell, 1995. **83**(5): p. 683-92.
2. Strawn, L.A., et al., *Minimal nuclear pore comlexes define FG repeat domains essential for transport.* Nat Cell Biol, 2004. **6**(3): p. 197-206.
3. Patel, S.S., et al., *Natively-unfolded nucleoporins gate protein diffusion across the nuclear pore complex.* Cell, 2007. **129**: p. in press.
4. Denning, D.P., et al., *Disorder in the nuclear pore complex: the FG repeat regions of nucleoporins are natively unfolded.* Proc Natl Acad Sci U S A, 2003. **100**(5): p. 2450-5.
5. Rout, M.P., et al., *Virtual gating and nuclear transport: the hole picture.* Trends Cell Biol, 2003. **13**(12): p. 622-8.
6. Lim, R.Y., et al., *Flexible phenylalanine-glycine nucleoporins as entropic barriers to nucleocytoplasmic transport.* Proc Natl Acad Sci U S A, 2006. **103**(25): p. 9512-7.
7. Ribbeck, K. and D. Gorlich, *The permeability barrier of nuclear pore complexes appears to operate via hydrophobic exclusion.* Embo J, 2002. **21**(11): p. 2664-71.
8. Baldwin, R.L. and B.H. Zimm, *Are denatured proteins ever random coils?* Proc Natl Acad Sci U S A, 2000. **97**(23): p. 12391-2.
9. Vucetic, S., et al., *Flavors of protein disorder.* Proteins, 2003. **52**(4): p. 573-84.
10. Bradley, P., K.M. Misura, and D. Baker, *Toward high-resolution de novo structure prediction for small proteins.* Science, 2005. **309**(5742): p. 1868-71.
11. Kuhlman, B., et al., *Design of a novel globular protein fold with atomic-level accuracy.* Science, 2003. **302**(5649): p. 1364-8.
12. Lockless, S.W. and R. Ranganathan, *Evolutionarily conserved pathways of energetic connectivity in protein families.* Science, 1999. **286**(5438): p. 295-9.
13. Schueler-Furman, O., et al., *Progress in modeling of protein structures and interactions.* Science, 2005. **310**(5748): p. 638-42.
14. Uversky, V.N., *Natively unfolded proteins: a point where biology waits for physics.* Protein Sci, 2002. **11**(4): p. 739-56.
15. Dyson, H.J. and P.E. Wright, *Intrinsically unstructured proteins and their functions.* Nat Rev Mol Cell Biol, 2005. **6**(3): p. 197-208.
16. Tompa, P., *The interplay between structure and function in intrinsically unstructured proteins.* FEBS Lett, 2005. **579**(15): p. 3346-54.
17. Allen, N.P., et al., *Deciphering networks of protein interactions at the nuclear pore complex.* Mol Cell Proteomics, 2002. **1**(12): p. 930-46.
18. Gilchrist, D., B. Mykytka, and M. Rexach, *Accelerating the rate of disassembly of karyopherin.cargo complexes.* J Biol Chem, 2002. **277**(20): p. 18161-72.
19. Jha, A.K., et al., *Statistical coil model of the unfolded state: resolving the reconciliation problem.* Proc Natl Acad Sci U S A, 2005. **102**(37): p. 13099-104.
20. Jha, A.K., et al., *Helix, sheet, and polyproline II frequencies and strong nearest neighbor effects in a restricted coil library.* Biochemistry, 2005. **44**(28): p. 9691-702.
21. Bernado, P., et al., *Defining long-range order and local disorder in native alpha-synuclein using residual dipolar couplings.* J Am Chem Soc, 2005. **127**(51): p. 17968-9.
22. Bernado, P., et al., *A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering.* Proc Natl Acad Sci U S A, 2005. **102**(47): p. 17002-7.
23. Tjandra, N. and A. Bax, *Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium.* Science, 1997. **278**(5340): p. 1111-4.

24. Ollerenshaw, J.E., et al., *Sparsely populated folding intermediates of the Fyn SH3 domain: matching native-centric essential dynamics and experiment.* Proc Natl Acad Sci U S A, 2004. **101**(41): p. 14748-53.
25. Wang, X., et al., *Characterizing the conformational ensemble of monomeric polyglutamine.* Proteins, 2006. **63**(2): p. 297-311.
26. Daura, X., et al., *The beta-peptide hairpin in solution: Conformational study of a beta-haxapeptide in methanol by NMR spectroscopy and MD simulation.* J. Am. Chem. Soc., 2001. **123**: p. 2393-2404.
27. Wèuthrich, K., *NMR in biological research : peptides and proteins*. 1976, Amsterdam, New York: North-Holland Pub. Co. ; sole distributors for the U.S.A. and Canada, American Elsevier Pub. Co. xii, 379 p.
28. Wèuthrich, K., *NMR of proteins and nucleic acids*. The George Fisher Baker non-resident lectureship in chemistry at Cornell University. 1986, New York: Wiley. xv, 292.
29. Bokor, M., et al., *NMR relaxation studies on the hydrate layer of intrinsically unstructured proteins.* Biophys J, 2005. **88**(3): p. 2030-7.
30. Csizmok, V., et al., *Primary contact sites in intrinsically unstructured proteins: the case of calpastatin and microtubule-associated protein 2.* Biochemistry, 2005. **44**(10): p. 3955-64.
31. Krishnan, V., *Determination of oligomeric state of proteins in solution from pulsed-field-gradient self-diffusion coefficient measurements. A comparison of experimental, theoretical, and hard-sphere approximated values.* JOURNAL OF MAGNETIC RESONANCE, 1997. **124**(2): p. 468-473.
32. Wilkins, D.K., et al., *Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques.* Biochemistry, 1999. **38**(50): p. 16424-31.
33. Tcherkasskaya, O., E. Davidson, and V. Uversky, *Biophysical Constraints for Protein Structure Prediction.* J Proteome Res, 2003. **2**(37-42).
34. Burley, S.K. and G.A. Petsko, *Aromatic-aromatic interaction: a mechanism of protein structure stabilization.* Science, 1985. **229**(4708): p. 23-8.
35. Callebaut, I., et al., *Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives.* Cell Mol Life Sci, 1997. **53**(8): p. 621-45.
36. Gaboriaud, C., et al., *Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences.* FEBS Lett, 1987. **224**(1): p. 149-55.
37. Dhe-Paganon, S., et al., *A phenylalanine zipper mediates APS dimerization.* Nat Struct Mol Biol, 2004. **11**(10): p. 968-74.
38. Maco, B., et al., *Nuclear pore complex structure and plasticity revealed by electron and atomic force microscopy.* Methods Mol Biol, 2006. **322**: p. 273-88.
39. Kustanovich, T. and Y. Rabin, *Metastable network model of protein transport through nuclear pores.* Biophys J, 2004. **86**(4): p. 2008-16.
40. Yang, Q., M.P. Rout, and C.W. Akey, *Three-dimensional architecture of the isolated yeast nuclear pore complex: functional and evolutionary implications.* Mol Cell, 1998. **2**: p. 223-234.
41. Wang, J., P. Cieplak, and P.A. Kollman, *How well does a RESP (restrained electrostatic potential) model do in calculating the conformational energies of organic and biological molecules?* Journal of Computational Chemistry, 2000. **21**: p. 1049-1074.
42. Case, D.A., et al., *The Amber biomolecular simulation programs.* J Comput Chem, 2005. **26**(16): p. 1668-88.
43. Wang, J., et al., *Development and testing of a general amber force field.* J Comput Chem, 2004. **25**(9): p. 1157-74.

44. Onufriev, A., D.A. Case, and D. Bashford, *Effective Born radii in the generalized Born approximation: the importance of being perfect.* J Comput Chem, 2002. **23**(14): p. 1297-304.
45. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.
46. Piotto, M., V. Saudek, and V. Sklenar, *Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions.* J Biomol NMR, 1992. **2**(6): p. 661-5.
47. Krishnan, V.V., K.H. Thornton, and M. Cosman, *An improved experimental scheme to measure self-diffusion coefficients of biomolecules with an advantageous use of radiation damping.* Chemical Physics Letters, 1999. **302**(3-4): p. 317-323.
48. Garcia de la Torre, J.G. and V.A. Bloomfield, *Hydrodynamic properties of complex, rigid, biological macromolecules: theory and applications.* Quarterly Reviews in Biophysics, 1981. **14**(1): p. 81-139.
49. Krishnan, V.V. and M. Cosman, *An empirical relationship between rotational correlation time and solvent accessible surface area.* Journal of Biomolecular NMR, 1998. **12**(1): p. 177-182.

Figure Legends:

Figure 1: (***A***) Diagram of the yeast nuclear pore complex and its GLFG nucleoporins in the transport conduit. The vertical tick marks in the FG nucleoporins mark the location of each FG motif: GLFG motifs are in yellow, FxFG motifs in red, and other variants in different colors. The indicated fragment of Nup116 (AA 348-458) was selected as a representative FG domain for this study. The F>A mutant version lacks the phenylalanines in FG motifs, which were replaced by alanines. The protein structures shown are a subset of the ten structures generated by MD simulations for each FG domain. (***B***) The AA sequence for the wild type and mutant Nup116 FG domains used. The phenylalanines (or alanines) in FG motifs are indicated in red. The AA sequences in gray are not part of Nup116, they were added to aid in the purification of the FG domain. The numbers in black bold font refer to the position of the indicated AA's in the Nup116 sequence; the numbers in gray refer to the position of the indicated AA in the FG domain protein analyzed.

Figure 2: Radius of gyration (Rh) as a function of simulation temperature.

Figure 3: (***A***) Histogram of end-to-end distances (calculated from the terminal C and N atoms) for 30,000 FG domain structures obtained by sampling every 0.5 ps the final 3 ns of each of the replicate ten MD simulations. (***B***) Histogram of radii of gyration (calculated using only the atoms in the peptide backbone) for 30,000 FG domains structures sampled as in A.

Figure 4: (***A***) Plots of the probability distribution of inter-atomic distances in the wild type and mutant FG domains. The distance distribution between phenylalanine (or alanine) residues in positions 32 and 71 is shown as a representative example. The dashed line in each plot corresponds to a Gaussian fit to the probability distribution. (***B***) Correlation plots between F-F or A-A pairs in the wild type and mutant Nup116 FG domains. The numbers in the axes correspond to the various F-F or AA pairs that result from all possible combinations (see Supplementary Table S2). The correlation map shows how each of the pairs is related to the others as derived using Pearson squared metric. (***C***) Schematic representations of the correlated Phe pairings in the wild type FG domain and the corresponding Ala pairings in the F>A mutant. The calculated average distance between Phe or Ala residues in the pair-wise

combinations is thickness coded; thick lines represent distances between 10 and 15 Å; medium lines represent distances between 15 and 20 Å; and thin lines represent distances greater than 20 Å.

Figure 5: (***A***) Aromatic and amide region of the water-gate residual water suppressed one-dimensional NMR spectra for the purified FG domains. Tall peaks in the spectrum between 7.1-7.3 ppm arise from the Phe residues in the wild type domain, which are absent in the F>A mutant. The sensitivity of the mutant FG domain spectrum is lower due to a lower protein concentration than wild type (***B***) Plot of the self-diffusion coefficient measurements performed using BPP-SED for the wild type and mutant FG domains. Circles depict the experimental points and squares and lines correspond to the fit to the diffusion data. The error bars are smaller than the size of the symbols.

Figure 6: Elution profile of the Nup116 FG domains in a sizing column. Purified wild type and F>A mutant Nup116 FG domains (100 µl of 7.5 mg/ml) were subjected to size-fractionation through a analytical-scale FPLC Superdex 75 column (24 ml volume) at a flow rate of 0.5 ml/min at 4°C. The column was pre-equilibrated in 20 mM Hepes pH 6.8, 150 mM KOAc, 2 mM MgAOc (***A***) or in 20 mM Hepes pH 6.8, 1 M KOAc, 2 mM MgAO, as indicated (***B***). The FG domain elution profiles were compared to those of carbonic anhydrase (29 kDa, Rs = 23.5Å), ovalbumin (45 kDa, Rs = 29.8Å) and BSA (68 kDa, Rs = 35.6Å), which were used as size standard. Note that the mutant FG domain elutes faster than the wild type version, indicating a less cohesive, more compact ensemble of structures.

Supplementary Figure Legends:

Figure S7**:** Rg values (Angstroms) from final 3 ns of MD simulation (at 300K) for the wild type and F>A mutant sampled every 1 ps. For clarity only 10 separate trajectories shown in each plot. The plots show no systematic increase or decrease in the Rg values over this simulation window.

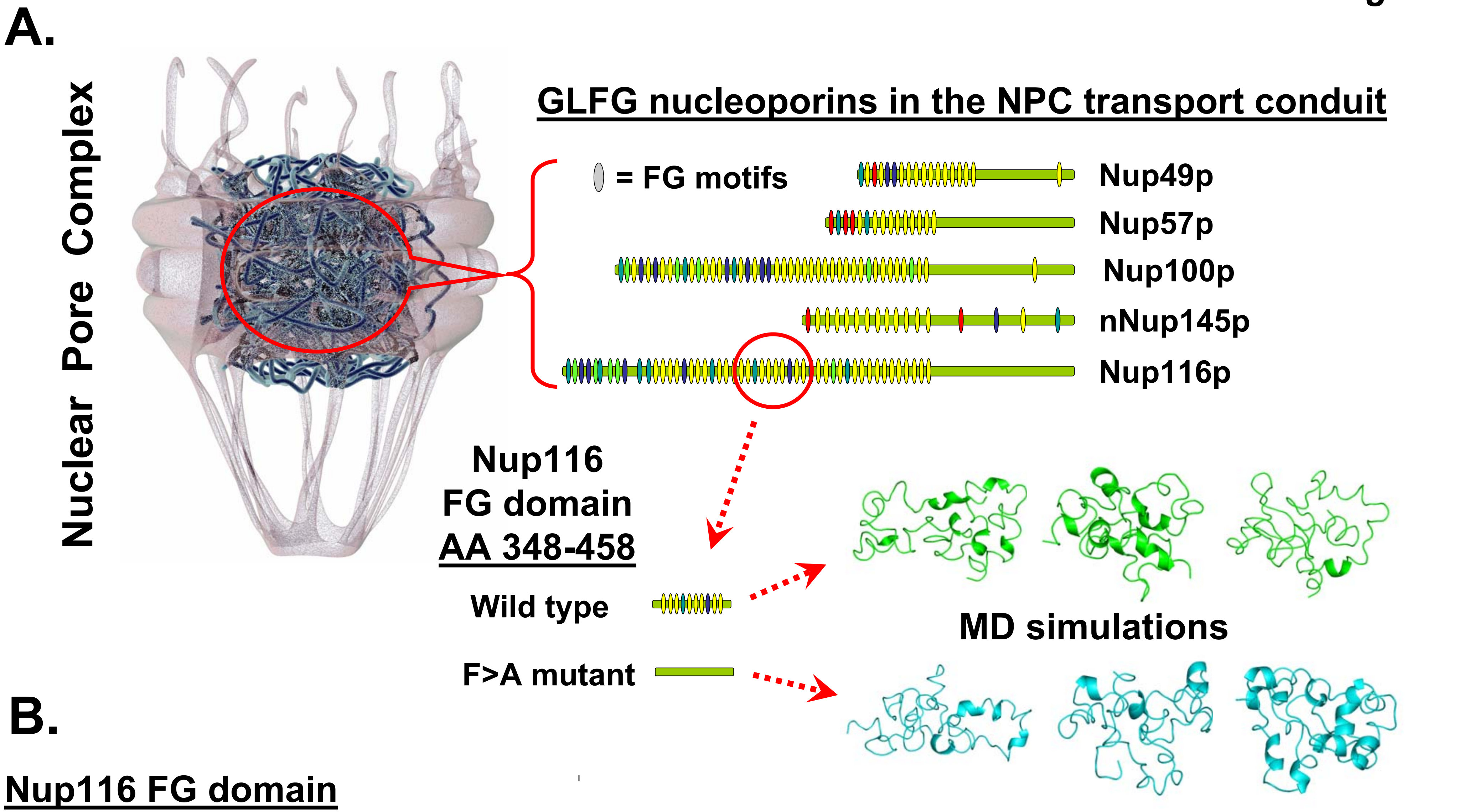
Figure S8: Autocorrelation of vector of all phi and psi angles calculated with a 1 ps time step averaging over all 2800 autocorrelation windows in the 3 ns simulations time. A separate line is

plotted for each of the twenty wild type replicates (blue lines), F>A mutant replicates (red lines), and the fibroblast growth factor as reference (green line).

Figure S9: Hydrophobic cluster analysis (HCA) of wild-type Nup 116 and F>A mutant. (a) HCA analysis of the wild-type sequence and F>A mutant was performed using the web server, http://bioserv.impmc.jussieu.fr/hca-file.html. The AA sequence is given in Figure 1 and the notation for the various symbols are defined in the original references on HCA [35, 36].

Krishnan et al, 2007

Figure 1



B.

Nup116 FG domain

Wild type

1 · 348 · 13 · 23 · 32 · 46 · 59

*GSRRASVGS*GAL**F**GAKPASGGL**F**GQSAGSKA**F**GMNTNPTGTTGGL**F**GQTNQQQSGGGL**F**GQQQ-
NSNAGGL**F**GQNNQSQNQSGL**F**GQQNSSNA**F**GQPQQQGGL**F**GSKPAGGL**F**GQQQGAST*HHHHHH*

71 · 84 · 93 · 103 · 112 · 458

F>A mutant

*GSRRASVGS*GAL**A**GAKPASGGL**A**GQSAGSKA**A**GMNTNPTGTTGGL**A**GQTNQQQSGGGL**A**GQQQ-
NSNAGGL**A**GQNNQSQNQSGL**A**GQQNSSNA**A**GQPQQQGGL**A**GSKPAGGL**A**GQQQGAST*HHHHHH*

Figure 2

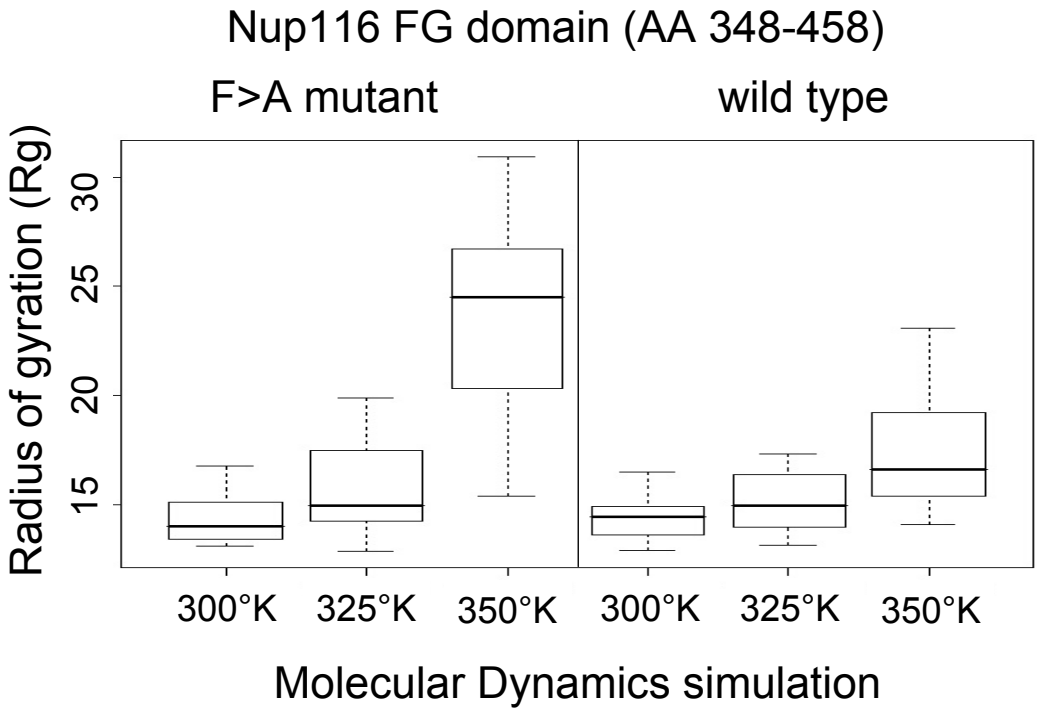
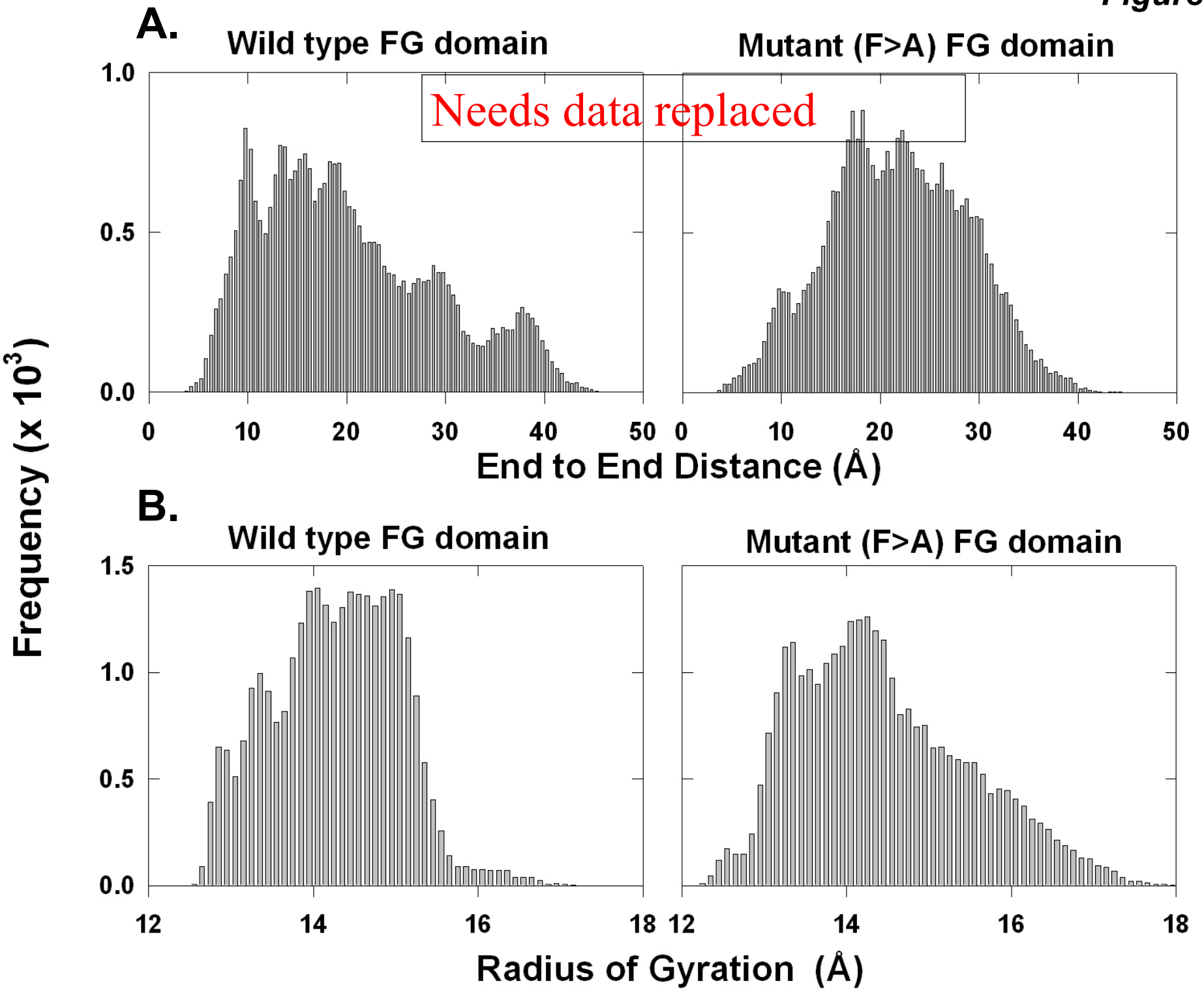
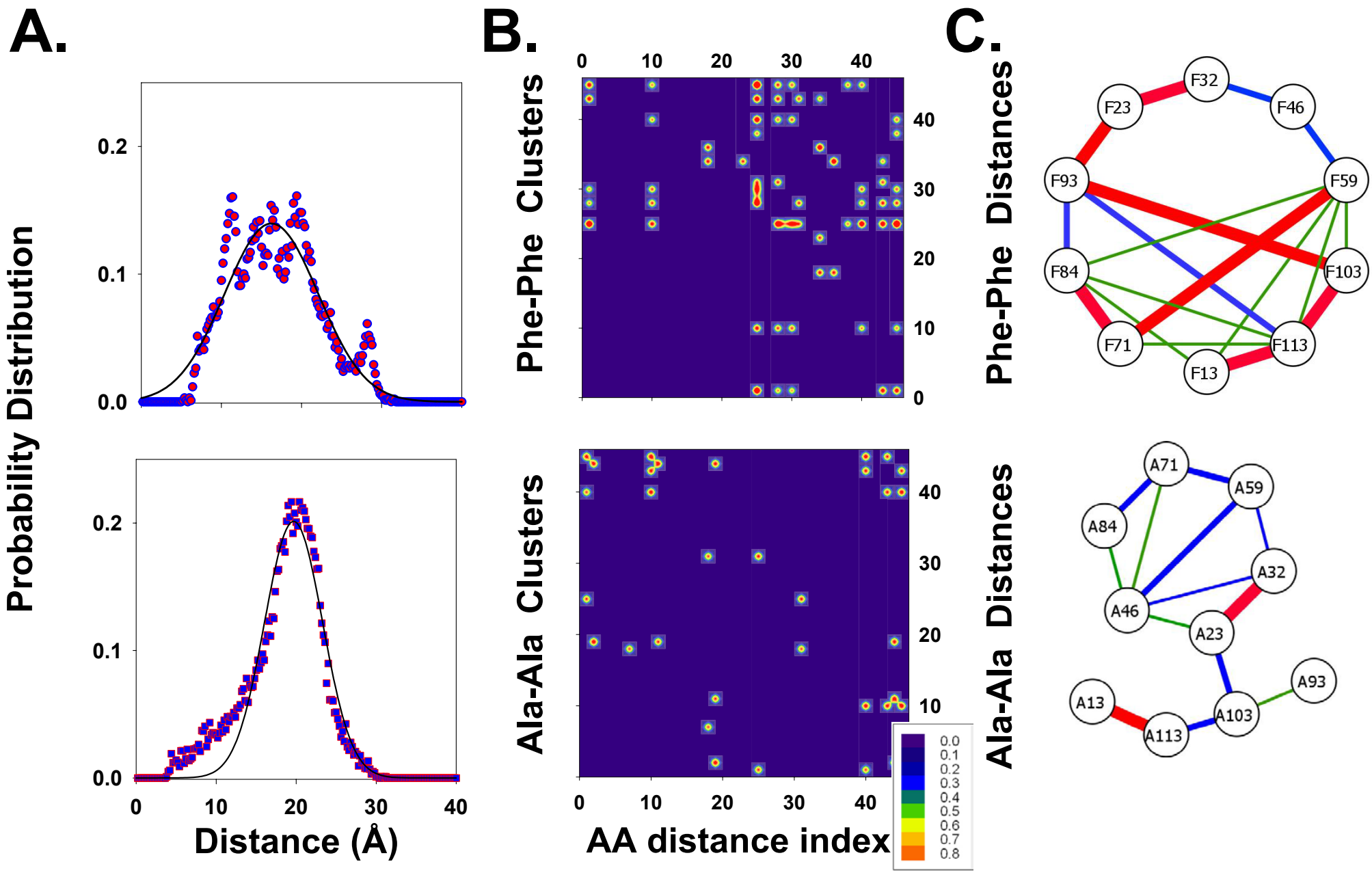

Nup116 FG domain (AA 348-458)
F>A mutant
wild type
Radius of gyration (Rg)
15
20
25
30
300°K
325°K
350°K
300°K
325°K
350°K
Molecular Dynamics simulation

Figure 3

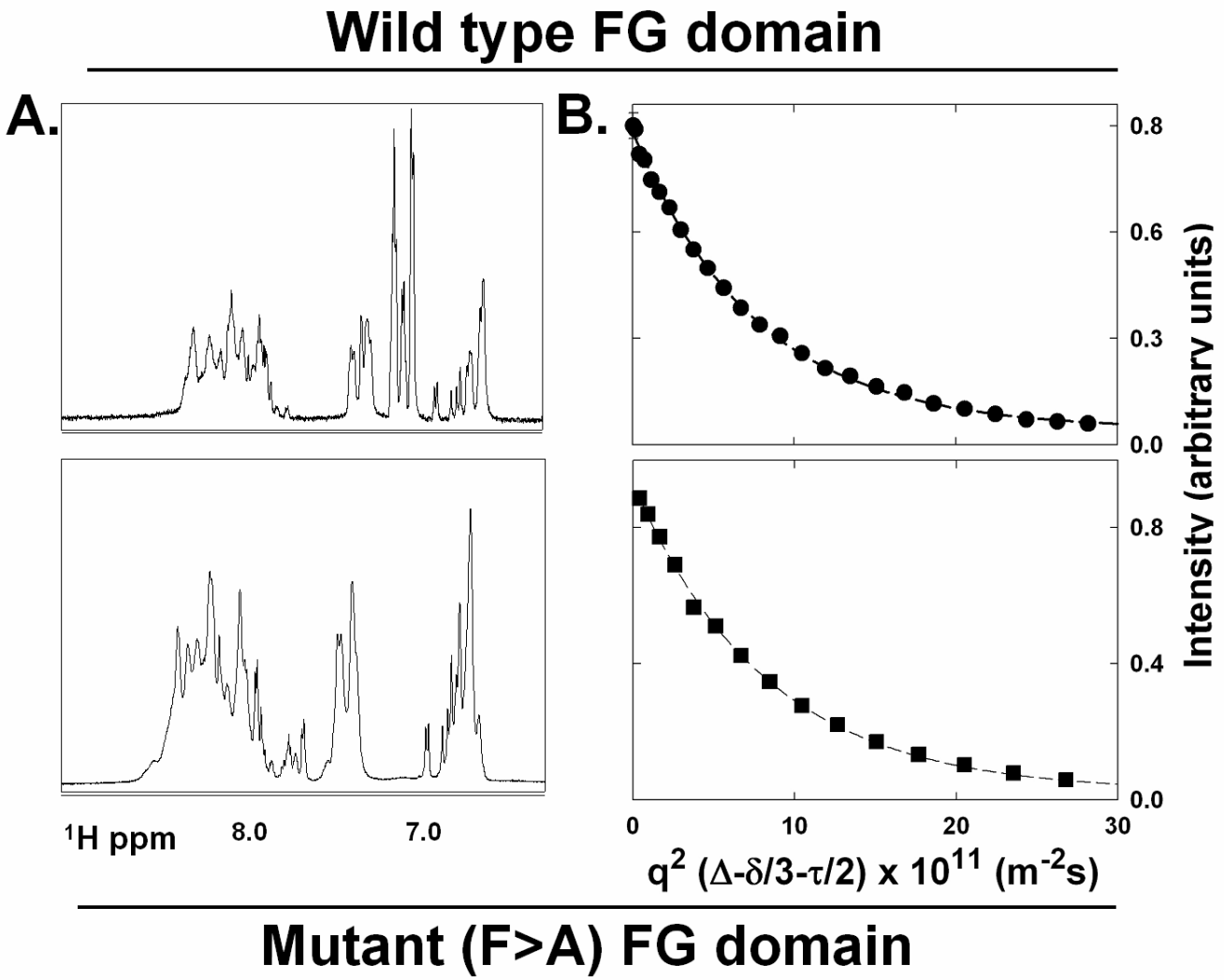

A.
Wild type FG domain
Mutant (F>A) FG domain
Needs data replaced
1.0
0.5
0.0
0
10
20
30
40
50
End to End Distance (Å)
B.
Wild type FG domain
Mutant (F>A) FG domain
1.5
1.0
0.5
0.0
12
14
16
18
Radius of Gyration (Å)
Frequency (x 10^3)

Wild type FG domain

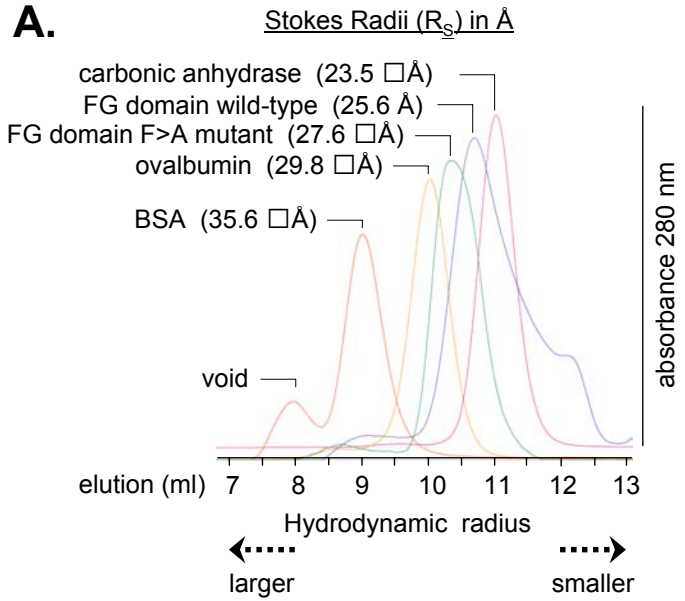


Mutant (F>A) FG domain

Figure 5


Wild type FG domain
A.
B.
0.8
0.6
0.3
0.0
0.8
0.4
0.0
Intensitv (arbitrarv units)
¹H ppm
8.0
7.0
0
10
20
30
q² (Δ-δ/3-τ/2) x 10¹¹ (m⁻²s)
Mutant (F>A) FG domain

A.



B.

	Superdex 75 elution volumes	
	column pre-equilibrated in buffer with	
	150 mM KOAc	1 M KOAc
carbonic anhydrase	10.998 ml	10.999 ml
ovalbumin	10.001 ml	10.004 ml
wild type FG domain	10.665 ml →	10.992 ml
F>A mutant FG domain	10.342 ml →	10.432 ml

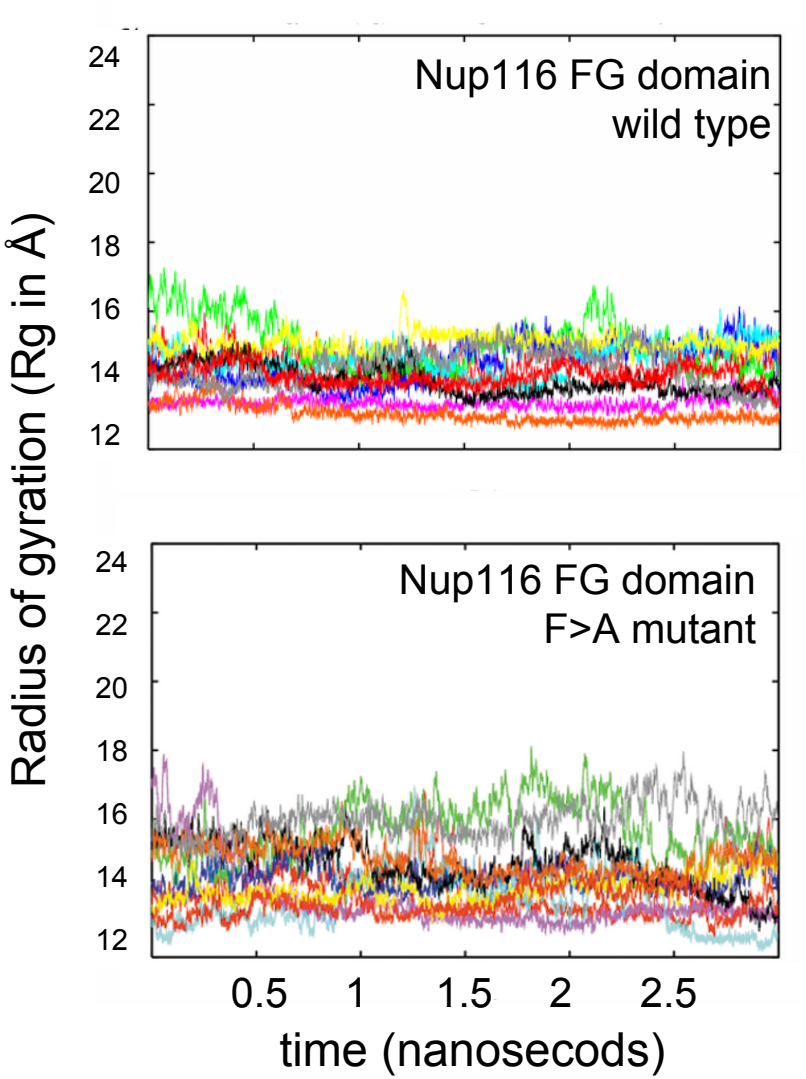
Table 1:

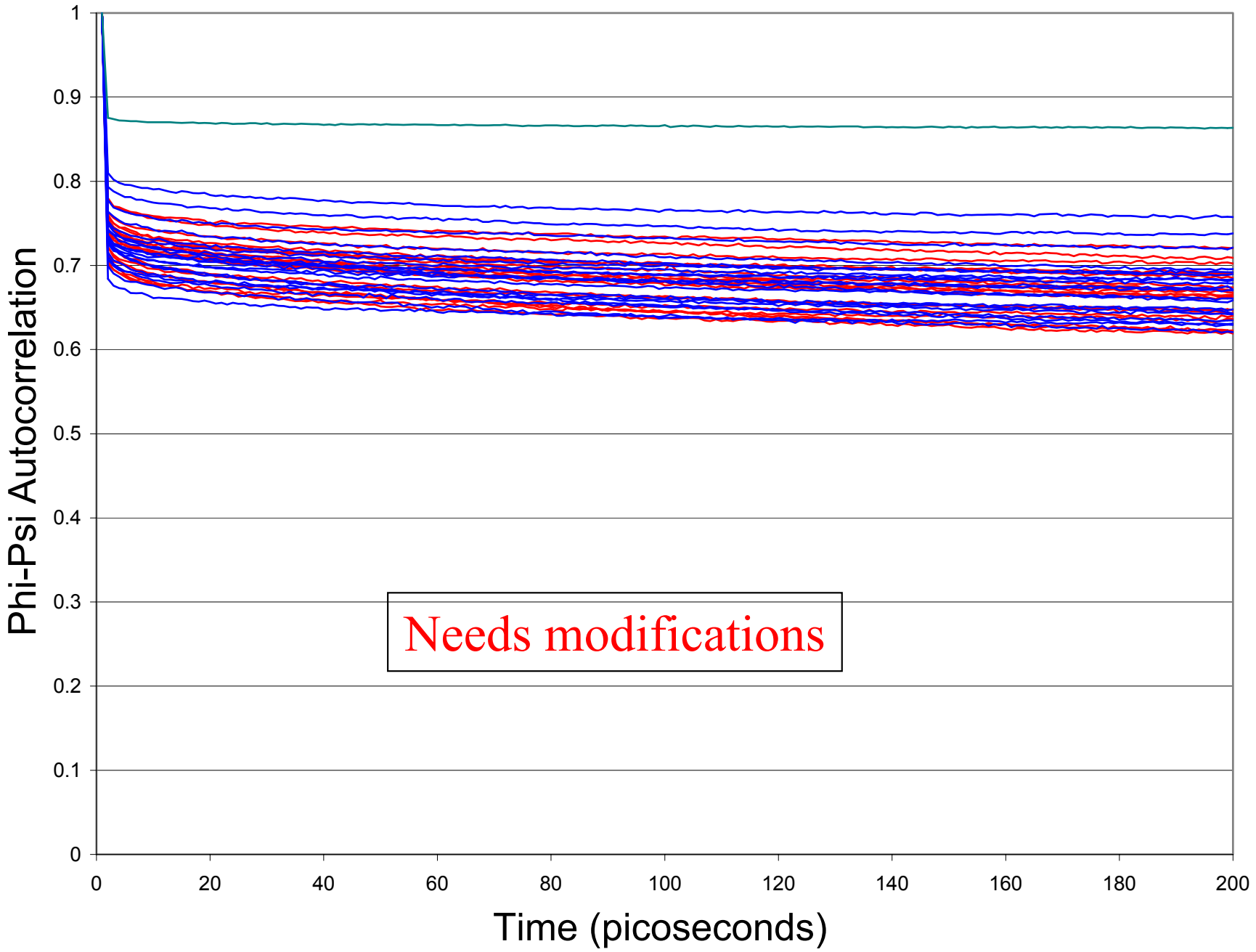
FG domain	AA's	MW (kDa)	sieving column measured Rs Stokes radius Rs (Å) = Rh	MD simulation predicted Rh at 350°K Rh (Å)	estimate from mass* predicted Rh (Å) native folded	molten globule	pre-molt globule	coil-like	urea	GdnHCl
Nup116 wild type	348-458	11.9 *	25.6**	22.5 ± 4.0*	16.6	19.9	25.6	27.8	29	30.2
Nup116 F>A mutant	348-458	11.1 *	27.6**	28.6 ± 5.2*	16.2	19.5	24.9	26.9	27.9	29.0

* In MD simulation it contains a 9 AA tag at the N-terminus in addition to the nup sequence

**contains a 9 AA N-terminal tag, and a 6xHIS tag at the C-terminus in addition to the nup sequence

The values highlighted in gray indicate best match to measured Rs.


Nup116 FG domain
wild type
Nup116 FG domain
F>A mutant
Radius of gyration (Rg in Å)
24
22
20
18
16
14
12
0.5
1
1.5
2
2.5
time (nanosecods)


Phi-Psi Autocorrelation
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
120
140
160
180
200
Time (picoseconds)
Needs modifications

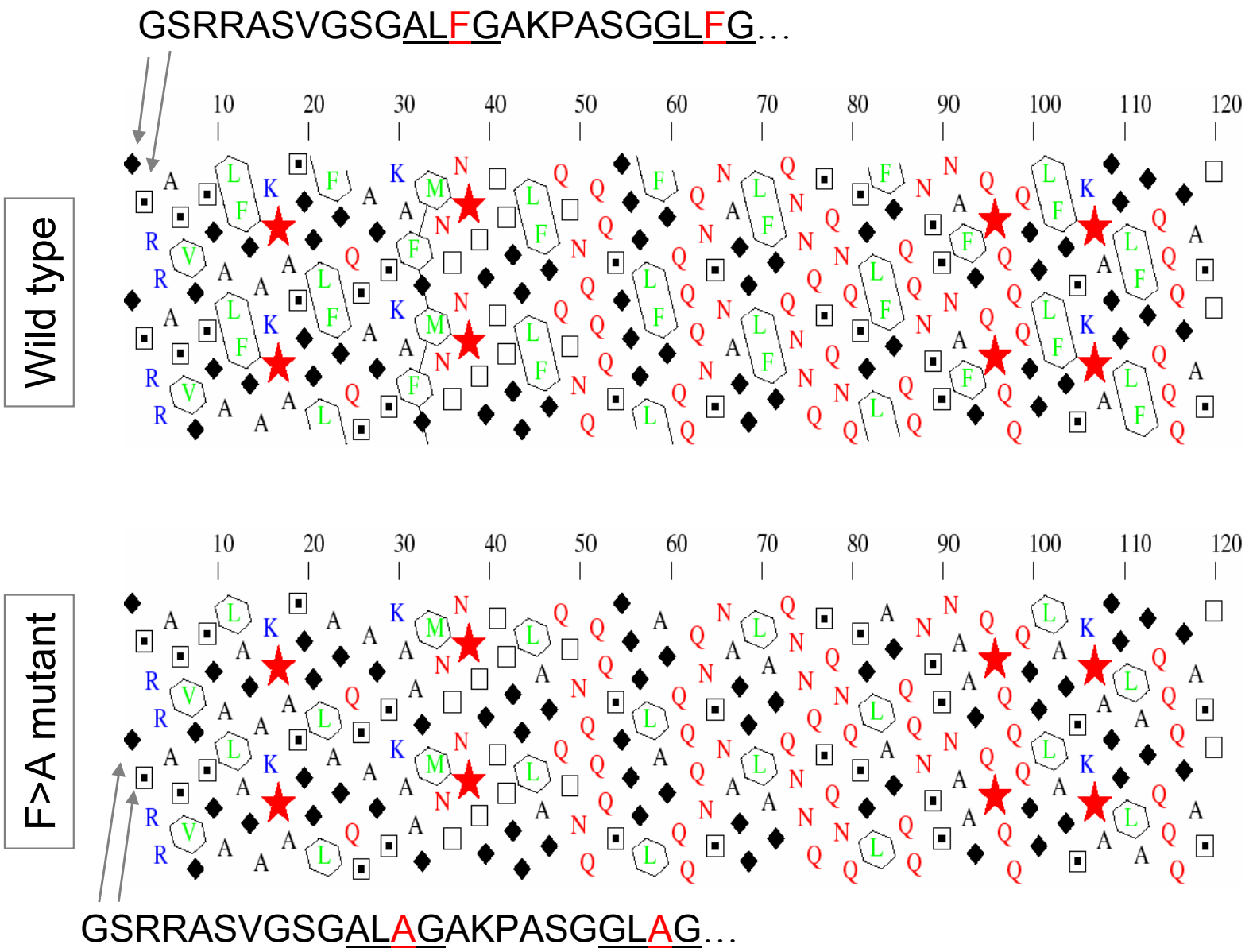

GSRRASVGSGALFGAKPASGGLFG…
Wild type
10
20
30
40
50
60
70
80
90
100
110
120
F>A mutant
GSRRASVGSGALAGAKPASGGLAG…

Table S2: List of 45 distance indices between the Cβ atoms of Phe or substitute Ala residues in the wild type or mutant FG domains, respectively.

index	Distance from		index	Distance from	
	From Cβ	To Cβ		From Cβ	To Cβ
1	13	46	24	13	46
2	59	71	25	59	71
3	59	84	26	59	84
4	59	93	27	59	93
5	59	103	28	59	103
6	59	113	29	59	113
7	13	59	30	13	59
8	71	84	31	71	84
9	71	93	32	71	93
10	71	103	33	71	103
11	71	113	34	71	113
12	13	71	35	13	71
13	84	93	36	84	93
14	84	103	37	84	103
15	84	113	38	84	113
16	13	84	39	13	84
17	93	103	40	93	103
18	93	113	41	93	113
19	13	93	42	13	93
20	103	113	43	103	113
21	13	103	44	13	103
22	13	113	45	13	113
23	46	113			

Table S3: List of distance constraints obtained from the Gaussian fit to inter-residue distance distribution.																
Gaussian fit results of distances (Å)									Gaussian fit results of distances (Å)							
	wild type FG domain				mutant (F>A) FG domain					wild type FG Domain				mutant (F>A) FG domain		
Index	Phe	Phe	Center	Width	Ala	Ala	Center	Width	Index	Phe	Phe	Center	Width	Ala	Ala	Center	Width
1	13	46	13.8	4.8	13	23	13.9	4.7	24	13	46	20.6	10.6	32	112	25.1	11.1
2	59	71	17.2	7.5	13	32	20.8	6.6	25	59	71	13.8	4.4	46	59	15.5	5.6
3	59	84	19.4	7.5	13	46	27.2	10.0	26	59	84	20.4	7.7	46	71	20.5	10.0
4	59	93	20.8	10.4	13	59	25.3	12.9	27	59	93	23.0	9.3	46	84	30.8	10.0
5	59	103	21.6	8.7	13	71	20.2	5.9	28	59	103	25.0	7.9	46	93	30.2	14.2
6	59	113	17.6	6.3	13	84	14.9	6.0	29	59	113	25.9	6.8	46	103	31.5	10.9
7	13	59	16.2	9.5	13	93	17.9	4.9	30	13	59	25.5	7.0	46	112	30.4	10.8
8	71	84	16.0	5.9	13	103	15.7	6.1	31	71	84	14.5	5.6	59	71	15.7	6.4
9	71	93	16.2	6.0	13	112	18.0	7.1	32	71	93	18.0	5.6	59	84	25.3	8.4
10	71	103	15.1	6.9	23	32	14.3	3.3	33	71	103	23.4	9.2	59	93	26.1	14.1
11	71	113	20.4	8.0	23	46	22.1	5.7	34	71	113	27.9	9.8	59	103	28.7	8.4
12	13	71	21.4	7.9	23	59	21.6	9.7	35	13	71	25.6	15.8	59	112	31.8	13.4
13	84	93	20.2	10.0	23	71	21.4	12.9	*36	84	93	16.3	10.2	71	84	19.7	7.1
14	84	103	20.5	9.6	23	84	24.1	13.3	37	84	103	24.1	8.8	71	93	22.1	9.8
15	84	113	20.8	10.5	23	93	21.9	11.5	38	84	113	25.2	9.1	71	103	26.9	9.6
16	13	84	19.8	8.3	23	103	18.4	5.7	39	13	84	24.4	10.4	71	112	26.2	9.6
17	93	103	20.6	7.4	23	112	21.9	11.3	40	93	103	14.2	4.7	84	93	24.4	4.3
18	93	113	16.0	7.0	32	46	17.4	5.0	41	93	113	18.4	9.9	84	103	19.9	8.8
19	13	93	16.5	7.3	32	59	21.5	6.8	42	13	93	20.5	9.3	84	112	22.9	9.4
20	103	113	20.8	9.3	32	71	23.8	11.2	43	103	113	14.3	5.7	93	103	15.3	3.6
21	13	103	19.7	7.0	32	84	28.3	12.6	44	13	103	21.0	7.2	93	112	21.7	4.6
22	13	113	22.3	7.7	32	93	26.1	16.4	45	13	113	13.7	3.6	103	112	13.8	3.9
23	46	113	24.0	12.2	32	103	23.7	13.9									

* Representative distance plot used in Figure 4.

FG nup		AAs	mass kDa	PMG volume nm3	number of copies per NPC	combined mass kDa	combined PMG volume nm3
nNup145	GLFG domain	1-219	21.6	144	16	346	2304
Nup57	GLFG domain	1-233	22.4	150	24	538	3600
Nup49	GLFG domain	1-246	23.8	162	24	571	3888
Nup100	GLFG domain	1-580	58.1	472	8	465	3776
Nup116	GLFG domain	1-696	69.5	585	16	1112	9360
NPC conduit volume (33 nm h x 25 nm r) = 64,821 nm3						3,032 kDa	22,928 nm3