Title

PHYSICAL MODELING OF GEOMETRICALLY CONFINED DISORDERED PROTEIN ASSEMBLIES

Author

Ando, David

Publication Date

2015

UNIVERSITY OF CALIFORNIA, MERCED

PHYSICAL MODELING OF GEOMETRICALLY CONFINED DISORDERED PROTEIN ASSEMBLIES

by

David Ando

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Physics

Committee in charge:
Professor Linda S. Hirst, Chair
Professor Michael E. Colvin
Professor Ajay Gopinathan, Advisor
Professor Michael Scheibner

2015

The dissertation of David Ando is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Professor Linda S. Hirst, Chair   Date

Professor Michael E. Colvin   Date

Professor Ajay Gopinathan, Advisor   Date

Professor Michael Scheibner   Date

University of California, Merced

To all those who toil to understand and protect, often under great hardship and sacrifice, the beauty we find in this world.

...... *We have begun to contemplate our origins: starstuff pondering the stars; organized assemblages of ten billion billion billion atoms considering the evolution of atoms; tracing the long journey by which, here at least, consciousness arose. Our loyalties are to the species and the planet. We speak for Earth. Our obligation to survive is owed not just to ourselves but also to that Cosmos, ancient and vast, from which we spring.* Carl Sagan, *Cosmos*

TABLE OF CONTENTS

Chapter

Appendix

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am forever indebted to my advisor Dr. Ajay Gopinathan, who has made my graduate studies in physics a most meaningful experience. Not only has almost everyday been one of discovery and learning, but I was given the freedom to learn and explore what was most important and interesting to me. I cannot overstate what a rare and profound opportunity this has been.

I would also like to thank the numerous faculty, research collaborators, and students who have provided tremendous support and guidance throughout my graduate studies. A big thank you to Jing Xu, Michael Colvin, Michael Rexach, Tim Connolly, Shawn Newsam, Linda Hirst, Michael Scheibner, Robby Puccinelli, Kostas Tsekouras, David Quint, Katie Copenhagen, KC Huang, Igor Goncharenko, Amanda Miguel, Nori Tani, and Nickolay Korabel.

I acknowledge support via internal funding from UC Merced, including the Physics Graduate Group, and a generous UC Merced Graduate Deans Dissertation Year Fellowship. I would also like to thank the National Science Foundation and the James S. McDonnell Foundation for extramural funding.

Chapter 3 is based on a manuscript previously published as Ando D *et al.* (2013) *Physical motif clustering within intrinsically disordered nucleoporin sequences reveals universal functional features.* PLOS One 8(9): e73831, used under CC BY with formatting changes and a streamlined introduction. I would like to acknowledge my co-authors Michael Colvin, Michael Rexach, and Ajay Gopinathan. Chapter 4 is based on a manuscript previously published as Ando D *et al.* (2014) *Nuclear Pore Complex Protein Sequences Determine Overall Copolymer Brush Structure and Function.* Biophysical journal 106(9): 1997-2007. This published work is reproduced under license 3663460362992. I would like to acknowledge my co-authors Roya Zandi, Yong Woon Kim, Michael Colvin, Michael Rexach, and Ajay Gopinathan. Some material from each of these papers has also been incorporated into Chapters 1 & 2.

CURRICULUM VITAE

David Ando
Ph.D. Candidate, Physics
University of California, Merced, School of Natural Sciences 5200 N. Lake Rd.
Merced, CA 95344
Advisor: Dr. Ajay Gopinathan

Education

- University of California, Merced, Merced, CA. Fall 2009 - current Ph.D. in Physics

- Chulalongkorn University, Bangkok, Thailand. Summer 2005 - Spring 2006 M.A. in International Economics and Finance, 2006

- University of California, Santa Barbara, Santa Barbara, CA. Fall 1998 - Summer 2002 B.A. in Physics and Mathematics, 2002

Publications

- Ando, D., and Gopinathan, A. "Cooperative interactions between different classes of disordered proteins play a functional role in the nuclear pore complex of Baker's yeast." submitted to *Physical Biology*, 2015

- Ando, D., Huang, K.C., and Gopinathan, A. "Cytoskeletal network morphology regulates intracellular transport dynamics." submitted to *Biophysical Journal*, 2015

- Ando, D., Xu, J., Gopinathan, A. "Cooperative filament switching emerges from inter-motor interference in multiple-motor transport." *Scientific Reports*, 4:7255, 2014.

- Ando, D., Zandi, R., Kim, Y.W., Colvin, M., Rexach, M., Gopinathan, A. “Nuclear Pore Complex Protein Sequences Determine Overall Copolymer Brush Structure and Function.” *Biophysical Journal*, 106(9):1997–2007, 2014.

- Ando, D., Colvin, M., Rexach, M., and Gopinathan, A. “Physical motif clustering within intrinsically disordered nucleoporin sequences reveals universal functional features.” *PloS one*, 8(9):e73831, 2013.

- Dow, S.F., Karcher, A., Levi, M., Momayezi, M., Von der Lippe, H., and Ando, D. “Design and Performance of the ELEFANT Digitizer IC for the BABAR Drift Chamber.” *IEEE Nuclear Science Symposium Conference Record*, 1:453–460, 1998.

Awards & Fellowships

- UC Merced Physics Graduate Group Annual Excellence Award for Outstanding Performance in Research and Publications 2015

- UC Merced Physics Graduate Group Annual Excellence Award for Outstanding Performance as a Teaching Assistant 2015

- UC Merced Graduate Deans Dissertation Year Fellowship 2014

- UC Merced Physics Department Travel Award 2012, ’13, ’14

- UC Santa Barbara RISE Scholarship in Condensed Matter Physics 2001

- Highest Honors, California State Examination in Chemistry 1998

Selected Presentations

- *Contributed poster,* Physics Meets Biology, Oxford, UK 2014

- *Invited talk,* TSRC Intrinsically Disordered Proteins: Sequence, Structure, Dynamics, and Function, Telluride, CO 2014

- *Contributed talk,* APS March Meeting, Baltimore, MD 2013

- *Contributed poster,* Biophysical Society Annual Meeting 2012, '13, '14

- *Contributed poster,* Bio Mechanical Engineering Conference, Stanford, CA 2011

ABSTRACT

Physical modeling of geometrically confined disordered protein assemblies
by
David Ando
Doctor of Philosophy in Physics
University Of California, Merced, 2015
Advisor: Professor Ajay Gopinathan
Committee Chair: Professor Linda S. Hirst

The transport of cargo across the nuclear membrane is highly selective and accomplished by a poorly understood mechanism involving hundreds of nucleoporins lining the inside of the nuclear pore complex (NPC). Currently, there is no clear picture of the overall structure formed by this collection of proteins within the pore, primarily due to their disordered nature and uncertainty regarding the properties of individual nucleoporins. We first study the defining characteristics of the amino acid sequences of nucleoporins through bioinformatics techniques, although bioinformatics of disordered proteins is especially challenging given high mutation rates for homologous proteins and that functionality may not be strongly related to sequence. Here we have performed a novel bioinformatic analysis, based on the spatial clustering of physically relevant features such as binding motifs and charges within disordered proteins, on thousands of FG motif containing nucleoporins (FG nups). The biophysical mechanism by which the critical FG nups regulate nucleocytoplasmic transport has remained elusive, yet our analysis revealed a set of highly conserved spatial features in the sequence structure of individual FG nups, such as the separation, localization, and ordering of FG motifs and charged residues along the protein chain. These sequence features are likely conserved due to a common functionality between species regarding how FG nups functionally regulate traffic, therefore these results constrain current models and eliminate proposed biophysical mechanisms responsible for regulation of nucleocytoplasmic traffic in the NPC which would not result in such a conserved amino acid sequence structure. Additionally, this method allows us to identify potentially functionally analogous disordered proteins across distantly related species.

To understand the physical implications of the sequence features on structure and dynamics of the nucleoporins, we performed coarse-grained simulations of nucleoporins to understand their individual polymer properties. Our results indicate

that different regions or blocks of an individual NPC protein can have distinctly different forms of disorder and that this property appears to be a conserved functional feature, consistent with the results of our physical bioinformatic analysis. Further simulations of grafted rings of FG nups mimicking the in vivo geometry of the NPC were performed and supplemented with polymer brush modeling to understand how aggregates of FG nups regulate transport in vivo. We found that the block structure at the individual protein level in terms of polymer properties is critical to the formation of a unique higher-order polymer brush architecture that can exist in distinct morphologies depending on the effective interaction energy between the phenylalanine glycine (FG) domains of different nups. Because the interactions between FG domains may be modulated by certain forms of transport factors, our results indicate that transitions between brush morphologies that correspond to open and closed states could play an important role in regulating transport across the NPC, suggesting novel forms of gated transport across membrane pores with wide biomimetic applicability in our *Diblock Copolymer Brush Gate* model.

Previous experimental research has concluded that FG nups from *S. cerevisiae* are present in a bimodal distribution, with the ”Forest Model” classifying FG nups as either diblock polymer like ”trees” or single block polymer like ”shrubs.” Our simulation and polymer brush modeling results indicated that the function of the tree FG nups in the *Diblock Copolymer Brush Gate* (DCBG) model is to form a higher-order polymer brush architecture which can open and close to regulate transport across the NPC. Here we perform coarse grained simulations of the shrub FG nups which confirm that they have a single block polymer structure rather than the diblock structure of tree nups. Our molecular simulations also demonstrate that these single block FG nups are likely compact collapsed coil polymers, implying that shrubs are generally localized to their grafting location within the NPC. We find that adding a layer of shrub FG nups to the DCBG model increases the range of cargo sizes which are able to translocate the pore through a cooperative effect involving shrub and tree nups. This effect can explain the puzzling connection between shrub FG nup deletion mutants in *S. cerevisiae* and the resulting failure of certain large cargo transport through the NPC. Facilitation of large cargo transport via single block and diblock FG nup cooperativity in the nuclear pore could provide a model mechanism for designing future biomimetic pores of greater applicability. In summary, this dissertation presents a cohesive body of research that uses a combination of techniques including bioinformatics, coarse grained molecular modeling, and polymer brush theory to understand the properties of individual FG nups and how they behave in aggregate, strongly constraining possible biophysical mechanisms which may play a role in regulating traffic through the NPC. Our results are observed across different species and are consistent with many experimental observations which have been reported. Finally, our DCBG model for NPC function provides testable predictions for future experimental investigation and provides a foundation for the design and commercialization of biomimetic pores for filtering applications *in vitro* and industrial use.

Chapter 1

INTRODUCTION

1.1 Nuclear Pores

Proteins are polymers of 20 different amino acid subunits, with each amino acid composed of a unique combination of carbon, hydrogen, sulfur, and nitrogen elements. Functional proteins in cells can fold into a rigid nano-mechanical structure or remain dynamic with significant disorder. These proteins which contain a disordered polymer like region are known as Intrinsically Disordered Proteins (IDPs). Folded proteins have been well studied for many decades via crystallographic techniques, which are able to elucidate their three dimensional structure. Function of these proteins is also strongly coupled to their relatively stable three dimensional structure. On the other hand, relatively little is known about IDPs as experimental and bioinformatic tools to probe their structure and function are in relative infancy. Although there is a spectrum of protein types from fully folded to completely disordered, proteins are typically defined as disordered if they contain at least one disordered region where backbone Ramachandran angles are dynamic. Disordered proteins are very common *in vivo* with, depending on the type of eukaryotic organism, roughly 36% to 63% of proteins containing at least one long disordered region greater than 40 amino acids in length [1]. Well studied IDPs include cellular tumor antigen p53 which functions to suppress tumor formation and therefore reduce the prevalence of cancer, Tau protein which helps to assemble the cellular cytoskeleton, and FG nucleoporins which help the Nuclear Pore Complex (NPC) to regulate transport between the cytoplasm and nucleus.

The functions of IDPs in the cell can be roughly divided into four broad categories: molecular recognition, molecular assembly/disassembly, protein modification and entropic chains [2]. IDPs whose function is to provide an entropic chain exhibit a range of rich behaviors that have great interest and application in polymer physics, statistical mechanics, and condensed matter physics. One such example where IDPs play an important role as entropic chains is in the NPC, which has a cylindrical geometry and is filled with IDPs called FG nucleoporins (Fig. 5.1). FG nucleoporins are commonly referred to as FG nups, and they are disordered NPC proteins which contain numerous phenylalanine glycine (FG) motif repeats. FG nups also function as disordered proteins for molecular recognition in the NPC, binding to transport factors to allow certain cargos through the pore. Unlike a folded protein which can perform molecular recognition under the lock and key paradigm given the relatively rigid structures of ligand and protein, IDPs can utilize their relatively high entropy to explore an ensemble of configurations for binding with a ligand.

NPCs efficiently regulate the movement of biomolecules between the nucleus of the cell and its cytoplasm, and are the only route of transport between the completely partitioned nucleus and cytoplasm. The efficient and comprehensive regulation of nucleocytoplasmic transport allows the NPC to protect the nucleus and to regulate the overall expressions of proteins within cells. Given the important role the NPC plays in eukaryotic cells, understanding its mechanism of function would be beneficial in understanding and treating malfunctions of NPC transport which lead to various diseases and cancers. Much of my research has focused on how these disordered FG nups can regulate traffic through the NPC.

Structurally the NPC is a supra-molecular structure composed of approximately thirty different types of FG nups [4]. A subset of the nups form a structural ring or cylindrical like structure which is embedded in the nuclear envelope. This structure forms an open aqueous channel, about 50 nm in diameter, which connects the nucleoplasm and cytoplasm (Fig. 5.1). A second subset of nups, those which form the focus of this dissertation, are grafted along the interior wall of the aqueous channel and are responsible for forming a selective diffusion barrier. Containing many phenylalanine-glycine (FG) repeats these nups are called FG nups [4] and are structurally unique due to their large FG repeat domains which are highly flexible and behave natively as polymer like unfolded proteins. Hence these FG nups are referred to as being intrinsically disordered or natively unfolded proteins [8]. FG repeats further separate into different types, with commonly studied GLFG and FxFG repeat motifs known to have different binding affinities [9]. In common baker's yeast *S. cerevisiae* there are ∼150 copies of FG nups in each NPC and it has been hypothesized that the NPC pore is occupied by dozens of FG nups that interact with each other weakly via hydrophobic attractions to form a network that functions as a semi-permeable diffusion barrier (see Fig. 5.1). This barrier maintains a tight seal against cytoplasmic particles larger than approximately 4 nm while allowing smaller particles to passively diffuse through. However, amazingly, it also allows the facilitated transport of specially tagged particles up to 40 nm diameter [10], at rapid speeds of 5-20 ms per transported cargo for even large mRNP complexes [11]. Additionally the NPC does not exert any forces or expend any energy, with the only known free energy consumption being limited to maintaining a gradient associated with the transport factors which generates the directionality of transport.

In vivo, traffic which an eukaryotic cell determines should translocate the NPC is tagged with a nuclear localization signal. In general, these signals are short amino acid sequences which "hang off" a cargo and tag it for transport. These nuclear localization signals then bind the cargo to NPC transport factor proteins, called importins. Finally, these transport factor proteins, while bound to cargo, will interact with the numerous FG motif repeats contained within disordered FG nups, with these FG nups lining the entire channel of the NPC, resulting in transport across the NPC. The precise physical mechanism of transport regulation which occurs once transport factors bind with FG motifs on FG nups has remained elusive.

Despite the NPC's key role in biology and numerous studies on the structure and properties of individual nucleoporins [4, 5, 8, 15–18, 28], the actual structure of the polymer complex within the nuclear pore and its mechanism of operation are still under debate. For example, prominent models of nuclear pore transport such as the "selective phase / hydrogel" [12] and "virtual gate / polymer brush" [14] models, assume very different morphologies for the polymer complex filling the nuclear pore (Fig. 5.1). The hydrogel model predicts that FG domains interact via hydrophobic amino acids to form a filamentous meshwork that physically blocks protein diffusion while the polymer brush model predicts that FG domains have limited hydrophobic interactions and instead behave as bristles that form an entropic gate at the NPC that blocks protein diffusion.

Most theoretical and polymer physics approaches [19–22] to this problem typically tend to assume a homogenous structure for individual FG nups, resulting in a relatively homogenous NPC architecture. There have also been recent attempts to perform coarse grained simulations of NPC proteins and the entire transport process through the NPC [23–26], but these too suffer from either the intrinsic drawback of assuming that the NPC proteins are homogeneous, or of over reliance on simple amino acid characterizations and assumptions, which could have left the ultimate underlying molecular architecture of the NPC unresolved. In this dissertation we attempt to overcome these issues by using a more reductionist and *ab initio* approach.

In Chapter 3, we use a comprehensive and novel bioinformatic analysis on all known disordered FG nups sequences. We show that the amino acid sequences of FG nups naturally separate into different regions or "blocks" (Fig. 1.2) and that biophysically important features within individual nups like the separation, spatial localization and ordering along the chain of FG and charge domains are highly conserved. Our current understanding of NPC structure and function needs to be revised to account for these common features that are functionally relevant for the underlying physical mechanism of NPC gating.

In Chapter 4, we use a detailed coarse grained (CG) molecular model that can reproduce the secondary and tertiary structure of proteins to study to the individual and aggregate properties of FG nups. This form of molecular modeling showed that different domains within an individual NPC protein can have distinct, quantifiably different forms of disorder and that these properties appear to also be a conserved functional feature. Many FG nups showed both extended coil conformations and collapsed coil conformations, spatially located along the FG nup amino acid sequence in a bimodal manner, Fig. 1.2. Contact analysis demonstrated that extended coil conformations were repulsive, that collapsed coil conformations were cohesive, and that extended coil conformations and collapsed coil conformations were repulsive with each other. The unique diblock polymer structure of the longer FG nups, that are critical for transport, under certain conditions within a confining cylindrical geometry, results in a mesoscale polymer brush structure with a dense FG domain core surrounded by a brush of disordered proteins with a high concentration of charged amino acids, see Fig. 1.3. To understand the physics behind this emergent structure we developed a simple polymer brush model based on the the properties of individual nups revealed by our CG model. Given that the interactions which hold the dense FG domain core together are likely be modulated by certain transport factors, our polymer brush modeling suggests that the approach of certain forms of cargo could alter the local brush structure, to allow for insertion and regulated transport.

In Chapter 5, we study the shortest FG nups which are commonly called "shrubs." Our CG modeling confirms that these proteins are collapsed coils, and we further extend our polymer brush model to include shrubs as compact FG domains which line the pore wall. In our revised brush model, the shrubs allow the pore to open to a greater extent, which is functionally beneficial for the transit of large cargos.

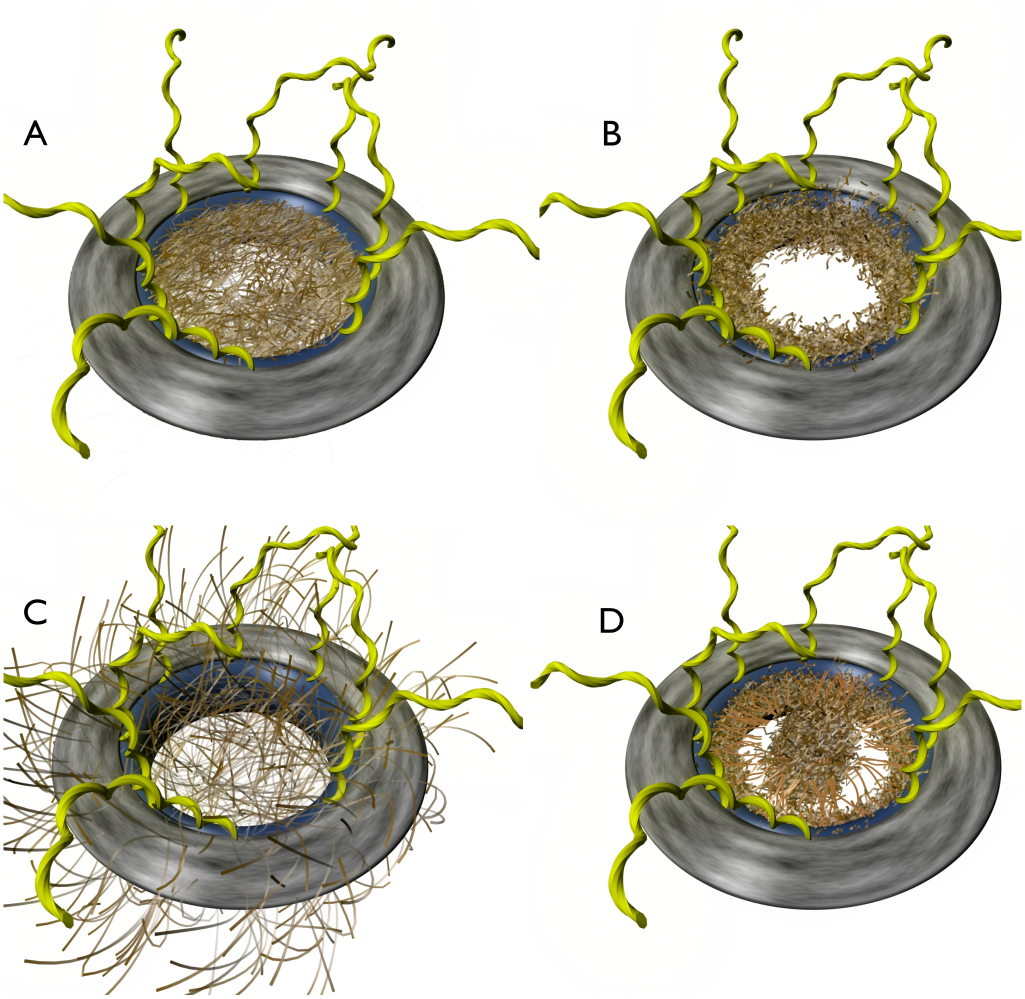

Figure 1.1: Illustration of NPC Models. Each NPC depicted spans the double lipid bilayer nuclear envelope, oriented such that the tops of the pores face the cytoplasm, while the bottoms of the pores face the nucleus. (A) FG nups form a hydrogel which transport factors and their cargo complexes transit by binding with FG motifs and temporarily disrupting cross links in the FG nup meshwork [12]. (B) FG Nups are collapsed and lie along the wall of the NPC, creating a surface to which transport factors bind and diffuse along in a dimensionally reduced manner [13]. (C) FG nups form a polymer brush which binds transport factors (possibly collapsing in their presence) but excludes unwanted molecules which don't interact with the FG nups [14]. (D) Individual FG nups can have collapsed coil gel-like regions and extended coil brush-like domains, resulting in a microphase separation of these domains within the NPC. A central plug-like structure and another dense shrub region which lies along the wall of the NPC are separated by a polymer brush of extended disordered regions (stalks) of FG nups. [3].

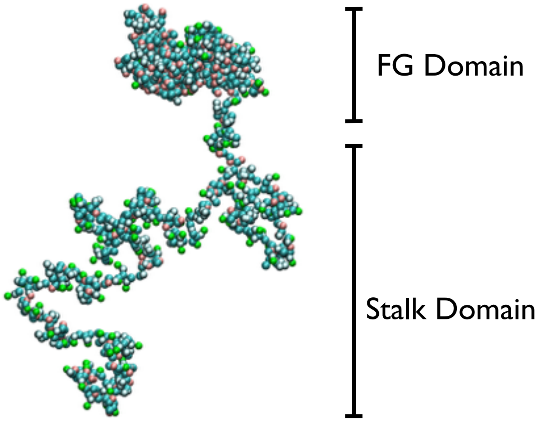

Figure 1.2: Depicted is a snapshot of Nsp1 from *S. cerevisiae* as seen in coarse grain molecular dynamics simulations. Many individual FG nups across different eukaryotic species have a biphasic structure, consisting of a collapsed block which is an "FG domain," rich in FG motifs while having a low number of charged amino acids, and an extended block "stalk domain," which has a low number of FG motifs while having a high density of charged amino acids.

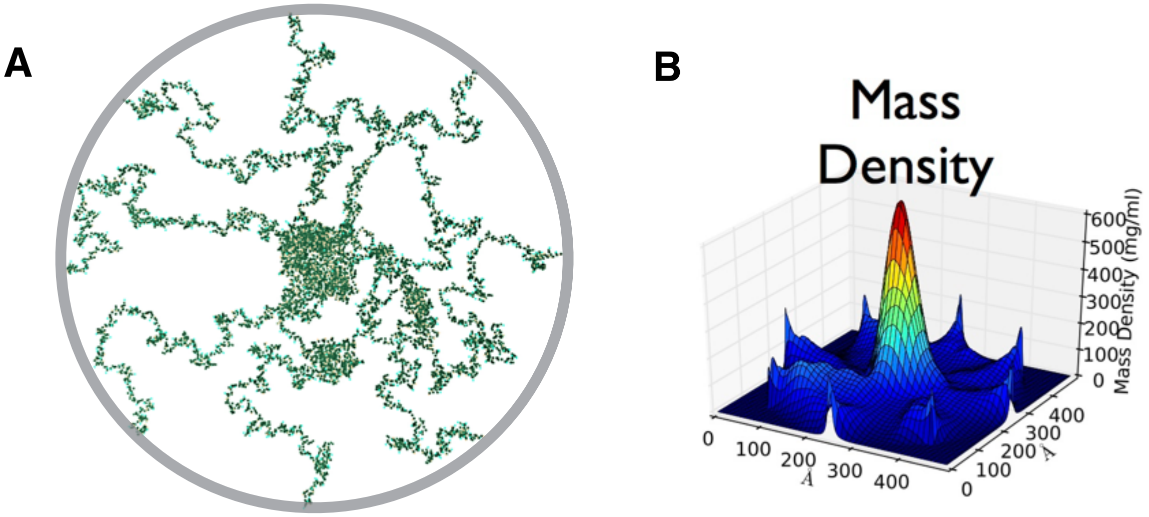

Figure 1.3: A) Snapshot of biphasic FG nups (Nsp1) confined to a cylinder similar in size to the NPC channel dimensions as seen in coarse grain molecular dynamics simulations. The tips of the disordered proteins can be seen to coalesce in the center of the pore supported by extend polymer brush like regions. B) Average mass density over time for these FG nups in the reflective boundary cylindrical pore. The total mass density clearly reveals a dense central plug connected by extended disordered stalk regions to the pore walls

Chapter 2

METHODS

2.1 Biohysical Amino Acid Sequence Analysis

In my research I have first focused on developing a novel form of bioinformatics for disordered proteins. This is especially challenging given the high mutation rates for disordered proteins and that functionality may not be strongly related to sequence. Bioinformatics in general is a scientific field focused on developing methods and software for the analysis and understanding of biological data. Much of bioinformatics is focused on understanding the amino acid sequences of proteins, which is especially challenging for disordered proteins such as FG nups. Bioinformatics of disordered proteins and FG nups is hard because the functionality of a protein may not be strongly related to the precise amino acid sequence. As these proteins do not fold, understanding their function through their amino acid sequence is difficult given that the sequence, to structure, to function paradigm that holds for folded proteins is not valid. Additionally, since disordered proteins do not fold the precise amino acid sequence is usually not important, resulting in high mutation rates, and little sequence similarity between homologous proteins. To overcome the difficulty that disordered protein functionality that may not be strongly related to sequence I have worked to develop a novel form of bioinformatic analysis, which can be applied to disordered proteins, that is based on the spatial clustering of physically relevant features such as binding motifs and charges along the amino acid chain. The insight behind this technique is that some disordered proteins may conserve certain spatial domains of general amino acid properties instead of a precise amino acid sequence for function. I have found this technique to be useful in finding a set of highly conserved spatial features in the sequence structure of individual FG nups, such as the separation, localization, and ordering of FG motifs and charged residues along the protein chain. These conserved features provide insight into the functioning of the pore and strongly constrain current models. Additionally this method allows for the identification of potentially functionally analogous disordered proteins across distantly related species in the face of essentially zero sequence similarity.

In our biophysical bioinformatic analysis of FG nups we first created to groups, the first group being FG nups and the second group a set of control proteins which are similar to FG nups. The first group, labeled NUP, contained proteins which were associated with the Nuclear Pore Complex via the keyword 'nucleoporin' in the Uniprot database (www.uniprot.org). Proteins in NUP which were classified by Uniprot as fragments rather than full length protein sequences were removed. A second group of proteins, labeled CONTROL, was created by systematically searching the Uniprot database for 12 proteins which contained 8 or more WG/GW motif repeats and which had substantial disorder, defined as containing disordered regions in excess of 100 amino acids in length as determined by the PONDR-FIT 2010 protein disorder prediction algorithm. Similarly 12 non-nucleoporins were taken from *S. cerevisiae* which contained FG motif repeats with greater than 100 disordered amino acids were taken from Nguyen et al [58] and added to CONTROL. Associated to each protein in these two groups was a collection of forty protein sequences, labeled ENSEMBLE, identical in all respects except for having the locations of all individual amino acids randomly permuted via a random shuffling.

Cluster identification was first done by transforming protein amino acid sequences into a series of binary representations with the specified motif set as unity and all other amino acids set to zero. The FG repeat containing proteins were parsed using the "FG" and "GF" motifs, WG repeats which were parsed using the "WG" and "GW" motifs, and charge motifs which were parsed together from the single AA motifs "D", "K", "E", and "R". The "GF" motif was used in addition to the "FG" motif due its chemical similarity and high degree of spatial correlation to "FG" motifs along the AA sequence of FG nups. The PreDeCon clustering algorithm was then applied to the one dimensional binary representations with a minimum cluster size of two amino acids (one AA for each cluster boundary) and a clustering resolution epsilon which maximized the Dunn validity index [53]. Clustering was done within the ELKI data mining software package [72]. A series of additional clustering variations were considered using various levels of reduced amino acid alphabets derived from the BLOSUM50 AA similarity [73].

All proteins which were not members of ENSEMBLE were additionally analyzed for the locations of disordered and folded regions using the PONDR-FIT 2010 algorithm [50]. A PONDR-FIT score for an amino acid less than 0.5 indicated a propensity for this location to be a folded region and in our analysis the protein was assumed to be fold at this location while a score greater than 0.5 indicated a disordered region and the protein was assumed to be natively unfolded at this location.

The NUP group proteins were characterized by the percentage of probable disorder and by their FG motif density (FG motif repeats/AA), where two groups arose naturally when restricting NUP to proteins with greater than 400 AAs, one with low percentage disorder and FG density, and another group with relatively high percentage disorder and FG density. We took proteins in the NUP group with greater than 0.15 FG/AA linear motif density and greater than 30% disorder to be the FG nups which we analyzed in this study.

For each FG/WG motif cluster in each protein, including the ENSEMBLE, the number of amino acids which were simultaneously in both charge clusters and the FG/WG motif cluster was counted and then divided by the number of total amino acids in the FG/WG cluster. This was the percentage overlap for clusters while the standard deviation of a nup in terms of cluster overlap measured the difference in total overlap between cluster types between a FG nup or control protein and its ENSEMBLE in standard deviations of the ENSEMBLE.

For each protein, including the ENSEMBLE, topological complexity was calculated by counting the number of times a sequence had a change in cluster type disregarding regions of the protein which did not fall within a FG/WG or charge cluster. Explicitly this was calculated by first computing a ϕ function mapping amino acid i to the values $\phi(i) = (InChargeCluster_i - InFGCluster_i)/(InChargeCluster_i + InFGCluster_i)$ for variables $In*_i$ true (1) or false (0) for AA i inside the indicated cluster type, omitting regions of the protein that had both $InChargeCluster_i$ and $InFGCluster_i$ equal to zero. The topological complexity was then derived by counting the number of times that this ϕ function had a non zero difference. In determining net charges of regions the only charged AAs were assumed to be D (-1), K (+1), E (-1), and R (+1) with Histidine excluded due to it being charged only 10% of the time at physiological ph [74] and because of its high depletion levels in FG nups [54]. All charge values are measured in normalized units of charge per AA length. The calculation of the net charge of folded regions of FG nups was restricted to folded regions with greater than 4 AA to avoid noise in the fold prediction algorithm. The net charge of disordered regions of FG nups which excludes FG clusters but contains charge clusters was determined by considering only the disordered regions which were mostly composed of charge clusters. With charge clusters having a charged AA density of roughly 30% we used a cutoff that required these disordered regions to have greater than 15% charged AAs.

The 1,167 FG nups analyzed were additionally clustered among themselves using the PreDeCon algorithm [51] over a five dimensional space spanning the properties measured in the paper. Setting the minimum cluster size to 50 proteins with a unitary neighborhood allowed for a large scale overview of how the 1,167 FG nups organize themselves in this five dimensional space, which resulted in four major groupings of FG nups.

2.2 Coarse Grained Molecular Dynamics

I used coarse grained molecular modeling of individual and aggregate FG nups to understand the dynamical and structural implications of the different types of disorder present within FG nups. Molecular dynamics in general is a simulation method *in silico* which uses classical approximations to the real quantum mechanical atomic interactions and dynamics found *in vivo*. This approach is typically used to model atoms bonded together to form molecules up to the size of single proteins and protein complexes, with simulation size limited by the computational complexity of this simulation technique. Coarse grained molecular dynamics simulations on the other hand average out many of the atomistic degrees of freedom, resulting in a much simpler molecular model which has a fundamental length scale larger than that of individual atoms. In single bead per amino acid residue type coarse grained models of proteins for example, each amino acid is treated as a single bead with a mass, charge, and simplified interaction potential which is representative of the full atomistic properties of the amino acid.

Starting from some initial condition where all coarse grained beads have a well defined position in space, coarse grained molecular dynamics simulates the dynamics of the particles through a time evolution process which integrates Newton's laws of motion. A commonly used time integration algorithm is the *velocity Verlet* scheme, where trajectories in the molecular dynamics simulation containing particle positions, velocities and accelerations at time $t + \Delta t$ given the state of the system at time t are derived from $\mathbf{r}$, $\mathbf{v}$, and $\mathbf{a}$ by the following equations:

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t)\Delta t + (1/2)\mathbf{a}(t)\Delta t^2 \tag{2.1}$$

$$\mathbf{v}(t + \Delta t/2) = \mathbf{v}(t) + (1/2)\mathbf{a}(t)\Delta t \tag{2.2}$$

$$\mathbf{a}(t + \Delta t) = -(1/m)\nabla V\left(\mathbf{r}(t + \Delta t)\right) \tag{2.3}$$

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t + \Delta t/2) + (1/2)\mathbf{a}(t + \Delta t)\Delta t \tag{2.4}$$

The software package used to perform coarse grained molecular dynamics in this dissertation is the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) software package [109], with the velocity verlet algorithm used for time integration. The coarse grained simulations performed do not use an explicit solvent to maximize simulation speed, but rather use the standard Langevin thermostat to model a background implicit solvent and recover Brownian dynamics of the simulation particles, while most solvent interactions with the simulated proteins are captured in the coarse grained interaction potential used from Hills *et al.* [27].

My coarse grain (CG) molecular dynamics simulations use the model developed by Hills *et al* [27], running in the LAMMPS software package in a computationally parallel mode over 24 cores in a dual Xeon server. Three separate CG models were created in an attempt to accurately simulate the behavior of the FG nups, with the different CG models producing disparate behavior to match different experiments. Each model, labeled α, β, γ had CG potentials scaled such that model results matched with the results of different experiments, consistent with how molecular CG models are commonly normalized [27, 77]. Scaling of CG models was achieved by multiplying the entire CG potential by a constant scaling factor.

The different CG models corresponded to different experiments on disordered FG nups as follows. CG model α matched radius of gyration measurements of human nup153 by Milles *et al* [78] using the FRET technique, which showed that nup153 resembles a compact collapsed coil disordered protien [79]. On the other hand, CG model γ was developed to match the experiments of Lim *et al* [80], which showed nup153 to be an extended coil disordered protein with a large radius of gyration, which only collapses in the presence of NPC transport factors. Finally, CG model γ was developed to reproduce the biphasicness of FG domain cohesion found by Yamada *et al* [3] through a series of bead-halo experiments. The biphasicness of FG nups from *S. cerevisiae* was found to involve the 'stalk' and 'FG' domains of the nups, where the 'stalk' domains are the regions of the FG nup with a low density of FG repeats and a high density of charged amino acids, while the 'FG' domain was the opposite, with a high density of FG repeats and a low density of charged amino acids. Stalk domains were found experimentally to be non-interacting, with no cohesion, while FG domains were cohesive to other FG domains. FG nups were also found to be biphasic in with respect to their radii of gyration, with stalk domains having expanded polymer conformations, while FG domains had more compact polymer conformations. CG model β was scaled to reproduce this measured biphasic behavior in terms of cohesion and radius of gyration.

Simulations were done at a temperature of 300 K, with an initial equilibration of the FG nups for 1 microsecond before production runs were started. The starting conditions used consisted of fully extended protein configurations for the simulations of the free FG nups, while FG nup starting conditions in the simulated pore consisted of FG nups placed bunched up against the pore wall near their grafting points. The nup specific sequences simulated were taken from Yamada *et al* [3]. During the initial 1 microsecond equilibration phase, all individual FG nups were able to fully collapse in under 300 ns. Production data was then taken from the following 4 microseconds of simulation after equilibration. Post simulation analysis included measuring the average radius of gyration of different domains of the FG nups and their average spatial densities and distributions. Trajectory snapshots were saved every 100 picoseconds during the 4 microseconds of production simulation. Contact proximity measurements between any two given simulated CG side-chain beads was measured by counting the number of times side-chains were within 16 Angstroms of each other in the saved trajectory, then normalizing by the number of trajectory frames.

2.3 Polymer Theory

Polymers are chains of monomers which form a flexible and dynamic 'string' like structure. Monomers may consist of molecules such as amino acids, and may vary widely in form within an individual polymer. The topology of polymers can be branched, forming for example a 'star' shape with several polymer chains connected to a central core monomer. In my research, I only consider linear non-branched polymers which have amino acids as monomers, the disordered protein FG nups of the NPC.

A surface which has a dense array of polymers grafted to one side forms a polymer brush. Polymer brushes are classified into two general types in relation to the grafting distance of the polymers, i.e. the average distance between polymer anchor points on the surface. The two types of polymer brush are 'mushroom' and 'true brush', which are limiting regimes that can be distinguished by either high or low grafting densities. At a low grafting density, individual polymers rarely interact with each other, forming an isolated mushroom shape whose size is roughly comparable to that of the radius of gyration free polymers. At high grafting densities, polymers interact with their neighbors and excluded volume interactions will push the polymers away from the grafting surface and increase the average height of the brush. This is the 'true brush' regime which we study in this paper for FG nups grafted to the interior of a cylindrical surface.

A simple analytical theory for polymer brushes was first proposed by Alexander [82], who devised an ingenuous method for solving for polymer brush properties as a function of individual polymer properties. His critical breakthrough was to assume that grafted polymers can each be individually thought of as a series of connected blobs, where inside each blob one can assume that the relation governing chain length and occupied polymer volume is the same as for a free polymer in the same solvent. Blobs are also assumed to have maximal size, where if they were made any larger, the relation governing chain length and occupied polymer volume for a free polymer would not hold. The cross sectional area of each blob was determined to be the area taken by individual polymers on the brush surface at their grafting points. The Alexander brush theory roughly captures the dependence and interaction between physical parameters of a brush and its constitute polymers such as polymer length, grafting density, and solvent quality on each other through a relatively accurate approximation for the calculation of the steric repulsion between blobs and the stretching energy of a polymer.

In the flat brush case, if one assumes that d is the average distance between polymer grafting sites and H the height of the uniform polymer brush then the volume of each chain is approximately d^2H. The free energy of excluded volume interactions in the brush can be thought of as excluded volume interactions between blobs in a chain, F_e, with N monomers of length a per chain. F_e is approximately the energy of blob overlap for each blob (kT) times the blob occupation probability for a given blob volume for every blob in each chain:

$$F_e = (kT) * ((\xi^3 c) * n_b) \tag{2.5}$$

with n_b the number of blobs per chain, ξ the radius of each blob, and c the concentration of blobs.

The free energy of the stretching of the polymers in the brush, F_s, is given by the reduction of degrees of freedom of the chain and the resulting reduction in free energy per chain of kT per blob formed as the polymer is stretched, times the number of blobs per chain n_b which are created:

$$F_s = (kT) * n_b \tag{2.6}$$

The number of blobs is simply:

$$n_b = H/\xi \tag{2.7}$$

with ξ the radius of each blob.

As we have assumed that inside each blob the relation governing chain length and occupied polymer volume was the same as for a free polymer in the same solvent, we can use the standard scaling relation that $N = n_b(\xi/a)^{5/3}$ for excluded volume chains [82], which together with equation 2.7 can be used to solve for $\xi = (N/H)^{3/2}a^{5/2}$. The free energy of the stretching of the polymers is therefore:

$$F_s = kT\frac{H^{5/2}}{a^{5/2}N^{3/2}} \tag{2.8}$$

We can now solve for c, the blob concentration, which geometrically is the number of blobs per chain ($n_b = H/\xi$) divided by the chain volume. Substitution of the relevant variables results in $c = (\frac{H}{aN^{3/5}})^{5/2}/(d^2H)$.

The total free energy F of the chains in the brush is finally the sum of the stretching and chain overlap free energies:

$$F = F_e + F_s = kT\frac{N^{3/2}a^{5/2}}{H^{1/2}d^2} + kT\frac{H^{5/2}}{a^{5/2}N^{3/2}} \tag{2.9}$$

We use this form of flat brush theory which has been adapted for use over negatively curved surfaces, such as for FG nups grafted to the inner surface of a cylinder, as solved in general by Sevick [88].

Chapter 3

FG NUCLEOPORIN BIOINFORMATICS

3.1 Biophysical Bioinformatics Introduction

Bioinformatic and proteomic analyses of the NPC have been successful in determining the structure, function, and origin of these structural elements of the pore [31, 36]. The disordered regions of nups have on the other hand generated much debate with regards to their underlying biophysical properties and function. Experimental results on FG nup aggregation and function are not all consistent [37–39], and traditional bioinformatics tools are unsuitable to predict function for these disordered proteins primarily due to their high AA substitution rates [40].

The reason for this is that typical bioinformatic sequence analysis involves pairwise comparison of amino acids (AAs) to reveal relationships among separate proteins [41], but this implicitly relies on a strong correlation between sequence, structure and function. However, since individual disordered proteins have the ability to form large functional ensembles of structures rather than a single folded structural state, they are much more robust to substantial changes in sequence over and above BLOSUM [42] or PAM [43] type similarity mutations, with often very-dissimilar sequences producing identical function [44]. In essence, it is not clear how disordered proteins retain functionality in the absence of a specific structure or a conserved sequence, both of which are normally associated with function in proteins [45].

It is therefore possible that the spatial distribution of physical properties along the disordered protein chain is a more relevant determinant of function than the specific AA sequence. Such an approach has been successful with other disordered protein systems such as the comprehensive analysis of 1,384 wild-type homeodomains by Vuzman *et al* [46]. They showed that the charge composition and distribution are evolutionary conserved, with positively charged amino acids forming dense and large clusters in order to functionally enhance binding to negatively charged DNA. This is a clear example that the sequence to structure to function paradigm well known among structured proteins can be applied at a coarse-grained level to intrinsically disordered proteins as well. Additionally, the general relationship between the sequence composition of disordered proteins, their structural characteristics, and function is currently a major area of research [47]. For example, Vucetic *et al* [48] have been able to broadly classify disordered protein sequences using a combination of sequence analysis tools into three groups arbitrarily labeled Flavor V, C, and S. Flavor V sequences tend to have positive charge and function as ribosomal proteins, Flavor C proteins tend to be neutral and function as modification sites, while Flavor S proteins tend to be negatively charged and function in protein binding.

Here we have taken the approach of analyzing the distribution of AA physico-chemical properties along FG nup polymers to model the copolymer block structure of these properties in order to generate metrics for comparison between FG nups that overcome highly divergent and noisy sequence data to reveal functional features. This approach also allows us to use these features to separate FG nups into distinct groups with potentially distinct functions, and to identify analogous subsets of these proteins among different species.

We assembled a data set for analysis of 3,355 nuclear pore related sequences (Supplementary Material) tagged with the keyword ”nucleoporin” from the Universal Protein Database [49]. This data set contained FG nups from 252 species (Supplementary Material). These proteins were characterized by the percentage of probable disorder as predicted by the PONDR algorithm [50] and by their FG motif distribution density (FG motifs/AA). Two groups arose naturally (Supplementary Material, Figure 3.6, Figure 3.7, Figure 3.8, Figure 3.9); one with low percentage disorder and FG density, which we identified as structural nups and kaps, and another group with relatively high percentage disorder and FG density, the FG nups. We focus on the second group for the rest of this analysis.

In this paper we use an approach called ”density-based clustering” where density is the number of occurrences of a specific motif (e.g. ”FG” AA pair or charged AA) per residue in a given segment of the protein. A cluster of a particular motif is defined as a segment of the protein with a higher density of that motif than the remainder of its amino acid sequence which models the copolymer block distribution for that motif, as depicted in Figure 3.1A. Note that individual occurrences of motifs appearing in low-density regions that separate clusters are considered to be noise in this clustering scheme. For this paper we adopted the PreDeCon density-clustering algorithm because it is determinate, robust against noise, efficient, and shows a superior clustering accuracy over other relevant methods [51]. We used a form of unsupervised clustering (see Methods) which determines the optimal clustering [52], via the Dunn index validity measure [53].

We first clustered two easily-defined, physico-chemical motifs within individual FG nups, the FG binding motifs and charged AAs, using our density clustering algorithm. We chose these two features because they are clearly defined and have been previously associated with FG nup shape and function [54, 55]. The FG sequence motif has been extensively studied and strongly implicated in transport factor binding [34, 35, 56], whereas the density of charged amino acids has been frequently noted as a predictor of intrinsically-disordered proteins and is generally an important predictor of polymer properties. For the sake of completeness, we also explored clustering across a wide spectrum of possible alternative motif choices (Supplementary Material, Figure 3.5), and concluded that the distribution of FG motifs and charged AA residues yielded the easiest to interpret results.

3.2 Results

3.2.1 Motif Clustering

We applied our clustering algorithm to the entire set of FG nups in all species studied to identify clusters that are enhanced in FG motifs and charged amino acids. FG clusters contained on average around 103 AAs, while charge clusters were smaller containing around 25 AAs on average. In analyzing these clusters we found that in all cases the FG nups have roughly 80% of their FG motif clusters located in regions which have absolutely no overlap with regions where charged AA clusters are found. This finding is shown in Figure 3.1B, which plots a histogram of the percent overlap in the FG and charged AA clusters (shown in blue). In order to examine the significance of this result for individual proteins, we generated a control ensemble of proteins consisting of random shufflings of each sequence (see Methods) and calculated the overlap of the FG motif and charged AA clusters in this ensemble; the results are plotted in Figure 3.1B (shown in red). For each of the FG nups we computed the number of standard deviations, σ that the overlap value is away from the mean overlap value of the shuffled ensemble (Figure 3.1B, inset, shown in blue). Approximately 90% of FG nups had an overlap that was one or more standard deviations less than their ensemble mean, and 77% had an overlap more than three standard deviations less than the ensemble average. In contrast the vast majority of FG clusters in the randomly-shuffled ensemble had large degrees of overlap with charge clusters, with only 4% having no overlap.

We also created a control group of 24 FG-nup-like proteins (Data S1), composed of WG motif argonaute binding proteins and non-nucleoporin FG repeat proteins, which share numerous similarities to FG nups including substantial amounts of disorder and a binding motif. Argonaute binding proteins contain WG motifs within their disordered domains, which are functional in substrate binding [57] similar to how FG motifs bind to karyopherins. An additional twelve non-nucleoporin FG-motif containing proteins involved in protein transport and sorting (similar to FG nups) were obtained from an exhaustive search of the *Saccharomyces cerevisiae* genome [58]. Notably, the control group exhibited little difference from the random ensemble in terms of FG (or WG) cluster overlap with the charged AA clusters, with greater than 85% of proteins having an overlap that was one standard deviation or less from the mean of the random ensemble (Figure 3.1B, inset, shown in red).

Although FG and charge clusters are highly disjoint, the ordering of these clusters along the protein chain can display various levels of spatial organization. To quantify this higher level cluster organization, we defined the topological complexity as the number of times that FG and charge clusters alternate with each other along the entire length of the amino acid polymer (see Methods for details). For example, a diblock polymer (i.e. a protein with a single cluster of charged AAs next to a single cluster of FG motifs) would alternate once between cluster types along the polymer resulting in a topological complexity of 1.

We found that over one third of FG nups adopt a simple diblock topological complexity of 1 (see Figure 3.1C, blue histogram), and another quarter of them adopt a topological complexity of 3, a quad-block structure (ABAB). In general, the distribution of the topological complexities of the FG nups is highly skewed, with 60% of FG nups having a topological complexity which is less than the overall mean by one or more standard deviations and 11% having a topological complexity greater than one standard deviation above their ensemble. In stark contrast to the FG nups is the control group which stands essentially indistinguishable from the random ensemble in terms of topological complexity, with more than 75% of proteins differing from the random ensemble by less than one standard deviation. The ordering of FG clusters relative to charged AA clusters is therefore very different than what would result from chance alone, suggesting a functional role.

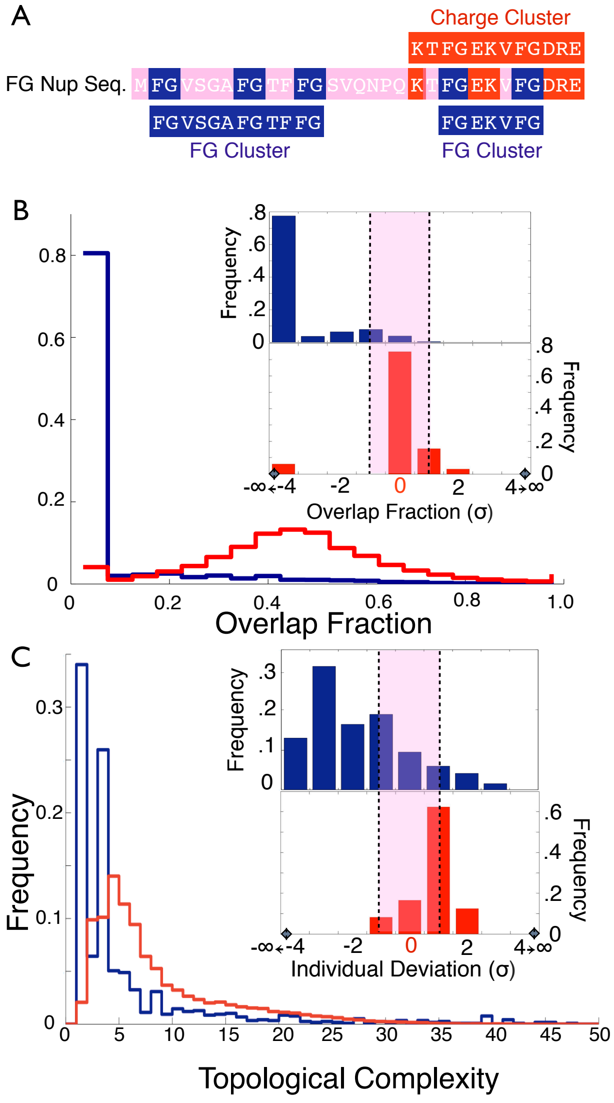

Figure 3.1: Spatial relationship between FG and charge clusters. (A) The FG clusters and charge clusters of an example FG nup sequence. (B) Percent overlap of FG clusters with charge clusters (blue). Almost 80 percent of FG motif cluster regions have zero overlap with clusters of charges. Other percentages of overlap, while strongly in the minority, appear with roughly equal probability. FG nups which have been randomly shuffled (in red) have 4 percent of their FG motif cluster regions completely disjoint from charge clusters while there is a strong tendency for FG clusters to overlap with charge clusters with a most probable frequency overlap at 45 percent.

Figure 3.1: Inset shows statistical significance of degree of overlap for FG nups (top, blue) versus WG/FG control group (bottom, red). Horizontal axis in inset shows the deviation of the percent overlap from the mean value for the randomly shuffled ensemble measured in units of the standard deviation of the random ensemble distribution. (B) Histogram of the topological complexity of FG nups (in blue) and randomly shuffled nups (in red). A majority of FG nups (66%) have a low topological complexity (less than 4) with 34% being purely diblock charge-FG copolymer structure, while randomly shuffled nups have only a small minority (22%) with low topological complexity and only 2% are diblocks. Upper inset shows that 77% of FG nups have a topological complexity which is less than their random ensemble by more than one standard deviation, while the control group shows little deviation from the ensemble (red, lower inset).

3.2.2 Relative Orientations

Given our result that the vast majority of FG nups have well separated charged AA and FG motif clusters, we explored the distribution of these clusters relative to the N- and C-termini. We defined the FG-charge polarity, η_{fc}, as the ratio of the distance between the centers of mass of FG and charged AA clusters to the total nup length, normalized such that $\eta_{fc} > 0$ indicates a charge to FG orientation coincident with the N-to-C terminus direction. This value ranges from 0 if for example the FG and charged AA motif clusters are wholly overlapping or interspaced, to values of ± 1 if the regions are disjoint and at opposite ends of the protein. Positive values of η_{fc} are therefore indicative of situations where the FG motifs cluster towards the C-terminus of the protein.

In the sequences analyzed, we found that FG nups have a strong tendency for $\eta_{fc} < 0$, with 70% having $\eta_{fc} < -0.35$. This indicates that FG clusters are preferentially located more towards the N-terminus of the protein than the charged AA clusters, and the centers of mass of the two motif clusters are separated by about half the protein length. Comparing to the results for the randomly-shuffled ensembles, we find 77% of FG nups having a lower (more negative) FG-charge polarity than the random ensemble mean by more than one standard deviation. In contrast, the FG-charge polarity computed for the control group of proteins resembles the random ensemble. Polarity for the randomly-shuffled ensembles of the FG nups showed no overall average polarity, as expected by symmetry.

The high conservation of the net negative FG-charge polarity ($\eta_{fc} < 0$) across a wide range of species suggests that this property is conserved by evolution and presumably related to NPC function. To relate this polarity to the spatial orientation of the FG nups in the NPC, we similarly looked at the polarity, η_{df}, between disordered and folded regions considering that the folded regions are predominantly anchor domains that connect the nups to the NPC scaffold [54].

The observed polarities were (Figure 3.2B) very similar to the average FG-charge cluster polarities, indicating that the clustering of the FG motifs is nearly always towards the N-terminal tip of the nups in disordered regions, away from an anchor domain which attaches to the NPC scaffold inner wall. Interestingly, FG-charge polarity is also found to be maintained locally within disordered regions for FG nups that contain at least one charge cluster in a disordered region, indicating a charge-containing spacer sequence (or entropic chain) closer to the wall, while a minority of FG nups had no charge clusters in their disordered regions (Figure 3.2B). This is consistent with the bimodal distribution of these features in nups from *S. cerevisiae* [54].

It is interesting to note that since mRNA translation into protein is initiated at the N-terminus, the observed disorder-folded polarity implies that the vast majority of nups have the disordered region synthesized prior to the ordered portion. This may have implications for the processing and transport of newly synthesized nups.

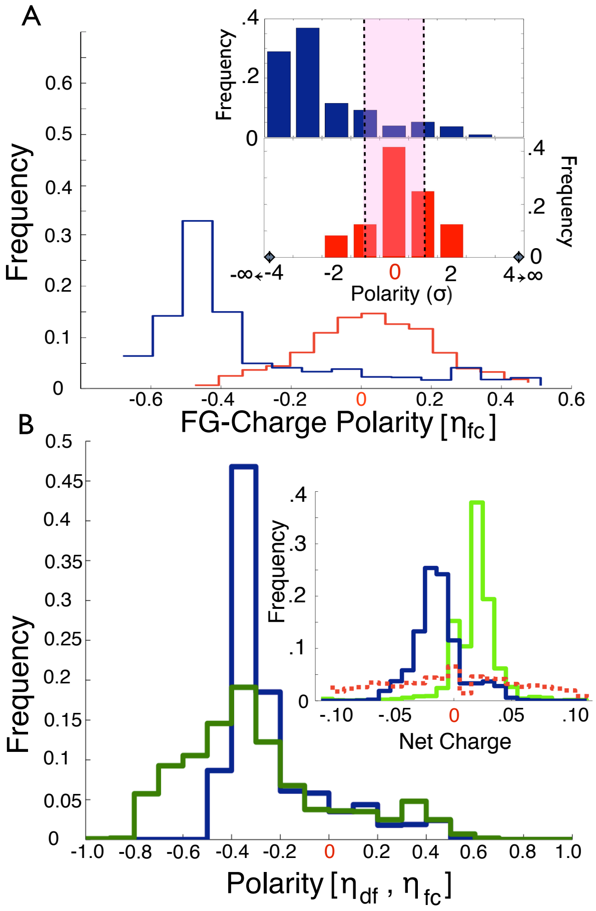

Figure 3.2: Polarity and charges of FG nups. (A) Polarity η_{fc} between charge and FG regions. FG nups (blue) tend to adopt a large negative and well conserved value for N-terminus to C-terminus polarity. Randomly shuffled FG nups showed no overall average polarity (red) and the statistical significance of FG nup polarity was consistently higher than three standard deviations (inset, blue, upper). The control group did not show a considerable difference from the random ensemble (inset, red, lower). (B) Polarity η_{df} between disordered and folded regions (blue) using the PONDR [50] protein disorder predictor. Observed polarities are on average similar to FG to charge cluster polarities, η_{fc}. Interestingly, η_{fc} values for the disordered regions alone (green) show a similar trend. Inset shows the net charge of folded structural nups/kaps (blue) and FG clusters (green) which appear to be equal and opposite on the whole, while disordered charge cluster regions (red dashed) appear to be net neutral. Histogram values for net charges for charge clusters greater than 0.1 e/AA and less than -0.1 e/AA are negligible and are shown in Supplementary Material.

3.2.3 Net Charges of FG Nup Regions

We next analyzed the exact distribution of charged AAs in FG nups and found that the net charge of FG clusters (Figure 3.2B, inset) is positive, and nearly equal and opposite to the net charge of kaps and structural nups. We also find that the disordered regions of FG nups, which exclude FG clusters but contain charge clusters (Supplementary Material, Figure 3.10), are neutral on average. The folded regions of FG nups on their own were also found to have zero net charge on average, but to associate with structural nups which are negatively-charged (Supplementary Material, Figure 3.11).

Previous results regarding net charges of nucleoporins by Ribbeck *et al* are consistent but less specific, showing that for two species, *S. cerevisiae* and humans, kaps are strongly biased towards having negative total charge [59], while the disordered regions of FG nups have a tendency to be positively-charged. We show this trend is common throughout all eukaryotic species, and more significantly that the positive charges are localized to the FG domains of N-terminal tips of disordered regions.

Figure 3.3A summarizes our findings by showing the effect on a FG nup sequence of successive constraints from the observed FG-charge overlap, cluster topology, charge and order-disorder polarities. The schematic at the top of Figure 3.3A shows the charged AA and FG motif patterns in a randomly shuffled AA sequence and the final schematic shows the typical FG nup that fits the observed sequence constraints. Figure 3.3B depicts the FG and charge clusters from *all* the FG nups from a single well characterized organism, *S. cerevisiae.* The representative FG nup depicted in Figure 3.3A is roughly similar to most of the FG nups present in *S. cerevisiae.* The deviations of individual nups from the representative distribution and from each other provide a means via which specific FG nups could adopt distinct functional roles within the NPC.

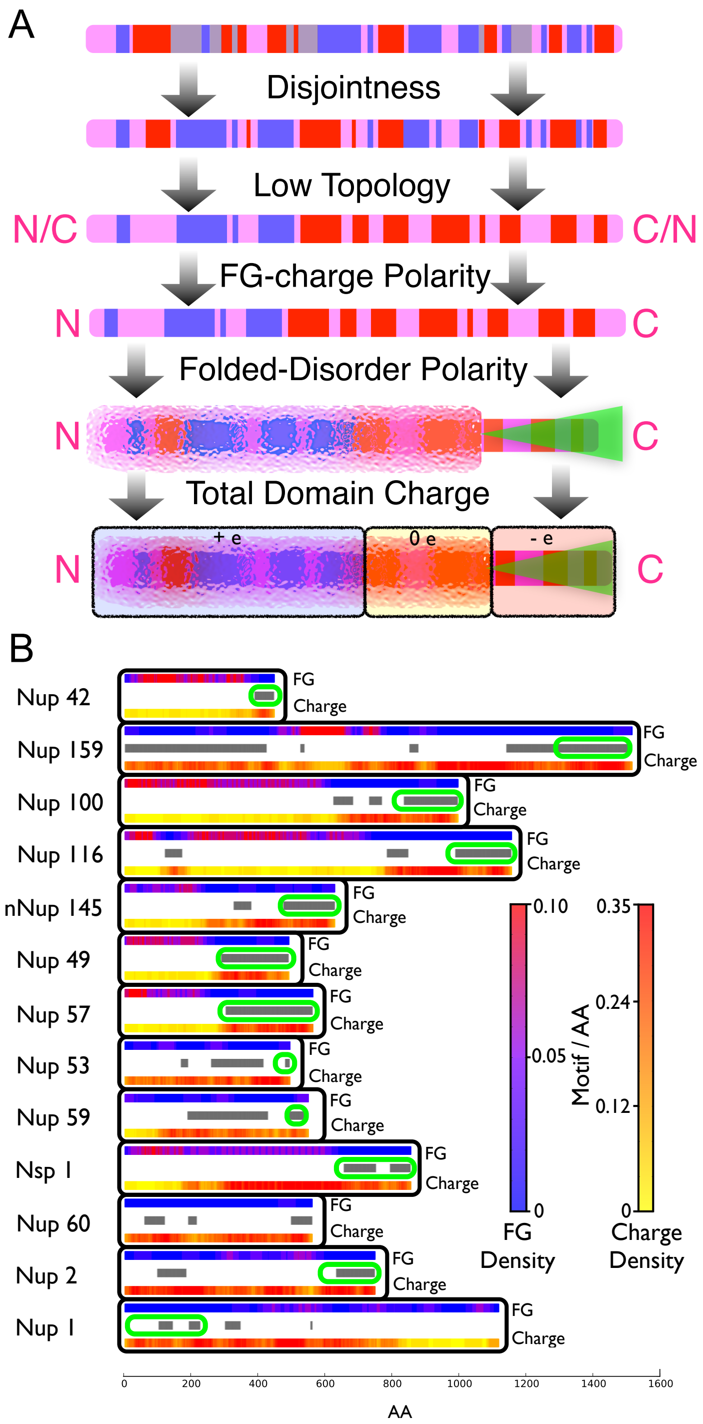

Figure 3.3: Consensus FG nup structure. (A) Effects of successively including observed sequence constraints. These constraints start from a typical randomly shuffled sequence at the very top with the pink bar representing the entire linear AA sequence.

Figure 3.3: Blue blocks represent FG clusters, red blocks represent clusters of charged AAs, gray represents overlap between cluster types, green shaded regions represent folded domains which anchor to the NPC wall, while disordered domains are represented by visually solvated pixelation. The arrows originating from the starting sequence represent a possible manner by which imposing the constraint of disjointness would result in a new sequence. Similarly the successive arrows represent the imposition of further constraints found in this paper, from the low topological complexity, to the FG-Charge polarity, to the Folded-Disordered polarity, to the net charge of domains, finally culminating in an average inference which is representative of FG nups. (B) The spatial distribution of FG motifs and charged AAs for all known FG nups of S. cerevisiae plotted as motif/AA, averaged over 20 nearest AAs. Regions of high FG motif density are shown in pink while regions of low charge density, also in pink, correspond spatially throughout the sequences of these nups. Regions of protein which are predicted to form folded structures by the PONDR algorithm are highlighted with grey bars, and known/predicted [44, 54] anchor domains circled with green ovals.

3.2.4 Functional Groupings of FG Nups

The fact that the spatial patterns of charged AAs and FG regions in FG nups which we have found, summarized in Figure 3.3A, are strongly conserved across eukarya provides evidence that these patterns are constrained by FG nup function. This suggests that the observed spatial patterns could provide templates for categorizing different FG nups into structural and functional subclasses. To test this, we constructed a five dimensional space determined by the biophysical metrics of FG-Charge cluster overlap, FG-Charge cluster polarity, Folded-Disordered region polarity, percent of disordered region composed of charged AA clusters, and topological complexity. Every FG nup analyzed in this paper is therefore a point in this space. We then used the density in this space to cluster these points into distinct groups. Clustering in this case is therefore not on motifs in an FG nup's amino acid sequence (as in the earlier part of the paper) but across sets of different proteins.

This clustering resulted in 4 distinct groups (Supplementary Material, Figure 3.12, Figure 3.13, Figure 3.14, Figure 3.15, Figure 3.16). The first group has a topological complexity of 1, a diblock structure, with a high level of disjointness and has no high density charged AA region in its disordered region. This group contains roughly one third of the FG nups with well-known FG nups in this group for *S. cerevisiae* being Nup57, and in humans Nup60 (Figure 3.4, Supplementary Material, Figure 3.17, Figure 3.18).

The second group is quite small at around 5% of FG nups and consists of those proteins with a topological complexity of 2, high disjointness and no other considerable properties. Given this group's small size and low occurrence among species, we concluded that proteins from this group are essentially outliers. On the other hand the third group is significantly sized containing approximately one quarter of FG nups with a topological complexity of 3. This group is notable for having large high density charged AA regions in its disordered regions, high disjointness and substantial negative polarities for both FG-charged AAs and folded to disordered regions. These are represented by FG nups such as Nup116 in *S. cerevisiae* and Nup98 in humans (Figure 3.4, Supplementary Material, Figure 3.19, Figure 3.20).

The fourth group which represents roughly one third of FG nups tends to have high degrees of topological complexity, large charged AA disordered regions, low disjointness, and low degrees of any polarity. Examples in this group include Nsp1 in *S. cerevisiae* and Nup153 in humans (Figure 3.4, Supplementary Material, Figure 3.21, Figure 3.22).

Based on the identity of S. cerevisiae proteins in each clustering group, these groups can be related to categories of FG nups previously identified in *S. cerevisiae* by Yamada *et al* [54]. Using a combination of biophysical and molecular simulation data they classified all S. cerevisiae FG nups as either compact structures (”shrubs”) or extended structures (”tree”), with the latter usually exhibiting separate regions enriched in charged AAs or FG motif sequences. Our clustering results find a similar subdivision, with cluster one corresponding to the ”shrubs”, and clusters three and four corresponding to different subclasses of ”trees”.

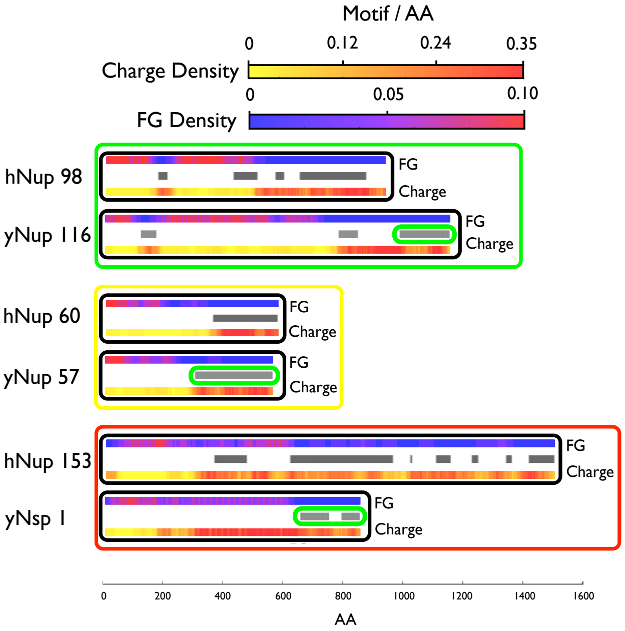

Figure 3.4: Identification of homologous proteins. Example proteins from *S. cerevisiae* ('y' prefix) and humans ('h' prefix) for each functional clustering group. Boxed in green are example proteins with low overlap between FG and charge clusters and a topological complexity of 3. In the yellow box are example proteins with low toplogical complexity and diblock structure, the first clustering group described in the text. Example proteins from the last clustering group described in the text are boxed in red at the bottom, with these proteins displaying high overlap and high levels of topological complexity (Nup153 sequence reversed to ease comparison with Nsp1).

3.3 Discussion

Regulation of nucleocytoplasmic transport by the NPC is governed by a permeability barrier whose structure, dynamics, and manner of interaction with transport factors remain elusive. Given that FG nups are a critical component of the NPC selective transport machinery [60], the conservation of observed FG motif and charge cluster arrangements over highly-divergent eukaryotic species suggest that the functional constraints of the NPC are leading to the observed patterns. These constraints could be at the level of the individual FG nups or to preserve essential interaction patterns between the FG nups.

At the individual protein level, any requirement for cooperative binding of FG motifs with transport factors (intra-strand cooperativity) could drive clustering but not necessarily the localization of FG motifs to the tips of nups. Inter-strand cooperativity between individual nups could, however, drive the localization of motifs to the tips of FG nups, because the cylindrical geometry [31] would require the tips to be in closer proximity (Figure 3.3B) [54]. Certain "fly casting" [39,61,62] mechanisms, where transport can potentially be driven by the action of single nups could also provide for a similar selective pressure at the individual FG nup level. Analogous mechanisms could be responsible for the evolutionary selection for spatial localization of positive charges to the tips of nups coincident with FG clusters, which would enhance binding with negatively-charged kaps. Depending on the geometric configuration this enhancement could be substantially above the levels proposed in [59] where charge localization was not considered.

At the collective level, the interactions between the large number of FG nups in the NPC have been hypothesized to lead to polymer brushes, hydrogels, reduced dimensionality surfaces, and other more complex structures, such as the transporter/plug structure. These emergent structures and their dynamics may require spatially-separated charge and FG clusters for numerous reasons. There have been a number of experimental observations which suggest FG nups interact to form larger scale emergent structures, including observations of a transporter structure [63], modern non-invasive FESE microscopy imaging of a central particle [64], AFM imaging of a central bump [65], and cargo specific spatial pathways through the NPC [66, 67].

Another element in FG nup sequence organization that offers clues to NPC function is that their charged disordered regions contain a large fraction of charged AAs (about 1/3 on average) yet are nearly neutral as a whole. This is unusual for disordered domains [68]. This selective pressure for net neutrality while containing a high density of charged residues, raises the possibility that electrostatic interactions could be a key part of the overall selective mechanism for NPC transport [69]. Current models of NPC transport [70, 71] typically involve homogeneous nups and are not based on any particular spatial distribution of FG motifs and charged residues, nor have any functional or structural need for entropic-chain domains with enhanced charge densities. Our findings provide a new set of evolutionarily conserved properties of the FG nups that need to be explained by proposed NPC transport models. More broadly, our approach of focusing on the abstract arrangement of physico-chemical features within sequences rather than the specific AA sequence shows promise in allowing for the identification of conserved features that share functionality across different types of proteins, but especially among functionally-equivalent disordered proteins with highly-disparate AA sequences.

In summary, our results argue that FG nup functionality is mediated via the arrangement of coarse-grained biophysical properties along the protein length, rather than by a precise sequence of specific AAs. This insight allowed us to identify templates that group FG nups into functional categories across widely different species despite their low sequence similarities. Moreover, the existence of these functional groups within the NPC of each species suggests a specialization of FG nups for different functional roles within the NPC. This approach can potentially be applied to gain functional and evolutionary insight into other classes of disordered proteins which have remained inaccessible to current bioinformatics techniques.

3.4 Methods

Two groups of proteins were created. The first group, labeled NUP, contained proteins which were associated with the Nuclear Pore Complex via the keyword 'nucleoporin' in the Uniprot database (www.uniprot.org) on June 29, 2012. Proteins in NUP which were classified as fragments rather than full length protein sequences were removed. A second group of proteins, labeled CONTROL, was created by systematically searching the Uniprot database for 12 proteins which contained 8 or more WG/GW motif repeats and which had substantial disorder, defined as containing disordered regions in excess of 100 amino acids in length as determined by the PONDR-FIT 2010 protein disorder prediction algorithm. Similarly 12 non-nucleoporins were taken from *S. cerevisiae* which contained FG motif repeats with greater than 100 disordered amino acids were taken from Nguyen et al [58] and added to CONTROL. Associated to each protein in these two groups was a collection of forty protein sequences, labeled ENSEMBLE, identical in all respects except for having the locations of all individual amino acids randomly permuted via a random shuffling.

Cluster identification was first done by transforming protein amino acid sequences into a series of binary representations with the specified motif set as unity and all other amino acids set to zero. The FG repeat containing proteins were parsed using the "FG" and "GF" motifs, WG repeats which were parsed using the "WG" and "GW" motifs, and charge motifs which were parsed together from the single AA motifs "D", "K", "E", and "R". The "GF" motif was used in addition to the "FG" motif due its chemical similarity and high degree of spatial correlation to "FG" motifs along the AA sequence of FG nups (Supplementary Material). The PreDeCon clustering algorithm was then applied to the one dimensional binary representations with a minimum cluster size of two amino acids (one AA for each cluster boundary) and a clustering resolution epsilon which maximized the Dunn validity index [53]. Clustering was done within the ELKI data mining software package [72]. A series of additional clustering variations were considered using various levels of reduced amino acid alphabets derived from the BLOSUM50 AA similarity [73] (Supplementary Material).

All proteins which were not members of ENSEMBLE were additionally analyzed for the locations of disordered and folded regions using the PONDR-FIT 2010 algorithm [50]. A PONDR-FIT score for an amino acid less than 0.5 indicated a propensity for this location to be a folded region and in our analysis the protein was assumed to be fold at this location while a score greater than 0.5 indicated a disordered region and the protein was assumed to be natively unfolded at this location.

The NUP group proteins were characterized by the percentage of probable disorder and by their FG motif density (FG motif repeats/AA), where two groups arose naturally when restricting NUP to proteins with greater than 400 AAs, one with low percentage disorder and FG density, and another group with relatively high percentage disorder and FG density (Supplementary Material). We took proteins in the NUP group with greater than 0.15 FG/AA linear motif density and greater than 30% disorder to be the FG nups which we analyzed in this study.

For each FG/WG motif cluster in each protein, including the ENSEMBLE, the number of amino acids which were simultaneously in both charge clusters and the FG/WG motif cluster was counted and then divided by the number of total amino acids in the FG/WG cluster. This was the percentage overlap for clusters while the standard deviation of a nup in terms of cluster overlap measured the difference in total overlap between cluster types between a FG nup or control protein and its ENSEMBLE in standard deviations of the ENSEMBLE.

For each protein, including the ENSEMBLE, topological complexity was calculated by counting the number of times a sequence had a change in cluster type disregarding regions of the protein which did not fall within a FG/WG or charge cluster. Explicitly this was calculated by first computing a ϕ function mapping amino acid i to the values $\phi(i) = (InChargeCluster_i - InFGCluster_i)/(InChargeCluster_i + InFGCluster_i)$ for variables $In{*}_i$ true (1) or false (0) for AA i inside the indicated cluster type, omitting regions of the protein that had both $InChargeCluster_i$ and $InFGCluster_i$ equal to zero. The topological complexity was then derived by counting the number of times that this ϕ function had a non zero difference. In determining net charges of regions the only charged AAs were assumed to be D (-1), K (+1), E (-1), and R (+1) with Histidine excluded due to it being charged only 10% of the time at physiological ph [74] and because of its high depletion levels in FG nups [54]. All charge values are measured in normalized units of charge per AA length. The calculation of the net charge of folded regions of FG nups was restricted to folded regions with greater than 4 AA to avoid noise in the fold prediction algorithm. The net charge of disordered regions of FG nups which excludes FG clusters but contains charge clusters was determined by considering only the disordered regions which were mostly composed of charge clusters. With charge clusters having a charged AA density of roughly 30% we used a cutoff that required these disordered regions to have greater than 15% charged AAs.

The 1,167 FG nups analyzed (Data S1) were additionally clustered among themselves using the PreDeCon algorithm [51] over a five dimensional space spanning the properties measured in the paper (Supplementary Material). Setting the minimum cluster size to 50 proteins with a unitary neighborhood allowed for a large scale overview of how the 1,167 FG nups organize themselves in this five dimensional space, which resulted in four major groupings of FG nups.

3.5 Supplementary Material

3.5.1 Exploration of different biophysical motif clusterings for FG nups

A series of clustering possibilities was considered starting with the hydrophobic/small amino acids (LVIMCAGSTPFYW) versus the hydrophilic amino acids (EDNQKRH) defined in the two letter reduced amino acid alphabet derived by grouping and averaging the full similarity matrix elements of BLOSUM50 [73]. This group showed considerable overlap between clustering types. The next case considered was the hydrophobic/small amino acids versus charged amino acids (EDKR), which performed in a similar manner. Considered next was the F modulo class of F equivalent AAs in the 4 letter BLOSUM derived alphabet versus the charged amino acids. This case was considered given the high enrichment of F amino acids in FG nups, and for the tendency of FG nups to have high density charged amino acids disordered regions which strongly affects polymer behavior. Possibly the most physically relevant property of FG nups is their FG motifs whose biochemical function is to bind transport receptors. This property was then clustered against the charged amino acids, which resulted in the most disjoint clustering class of all of the clustering groups tested, with 70% of FG clusters found to be more than 90% disjoint from charge clusters and nearly no FG clusters having greater than 90% overlap with charge clusters (Fig. S1). The next biophysical properties considered were simplified versions of FG motif vs. charged amino acids which yielded lower values of disjointness, similar to the results obtained with the other generalizations of these motifs considered. These other cases involved further simplified motif clustering possibilities including reducing in the FG motif to simply F as well as representing charges only by the most highly enriched charged amino acid in nups, K [54].

Explorations across possible biophysical clustering types was limited to proteins from distantly related Saccharomyces species and Homo sapiens FG nup proteins contained within the Uniprot (www.uniprot.org) database to conserve computational resources, while a full exploration of all nucleoporins was done only for FG and charged AAs. The 85 Saccharomyces and Homo genus proteins analyzed are listed in SI Data.

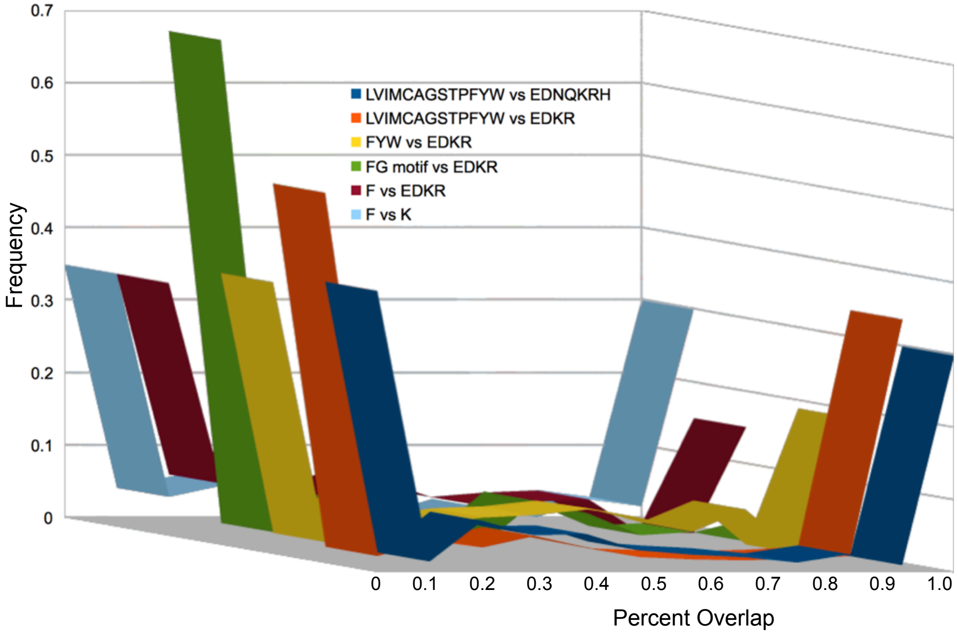

Figure 3.5: Histogram of percentage overlap for generalized and simplified FG motifs (LVIMCAGSTPFYW, FYW, and F) versus various characterizations of polar amino acids (EDNQKRH, EDKR, and K). Analyzing FG nups using FG motif versus charged AAs results in a clustering with the least degree of overlap, and is therefore the simplest and easiest to interpret.

3.5.2 Identification of FG nups

Initially all NUP proteins were analyzed for percent disorder and total FG density which resulted in a relatively continuous distribution as can be seen in Fig. S2.

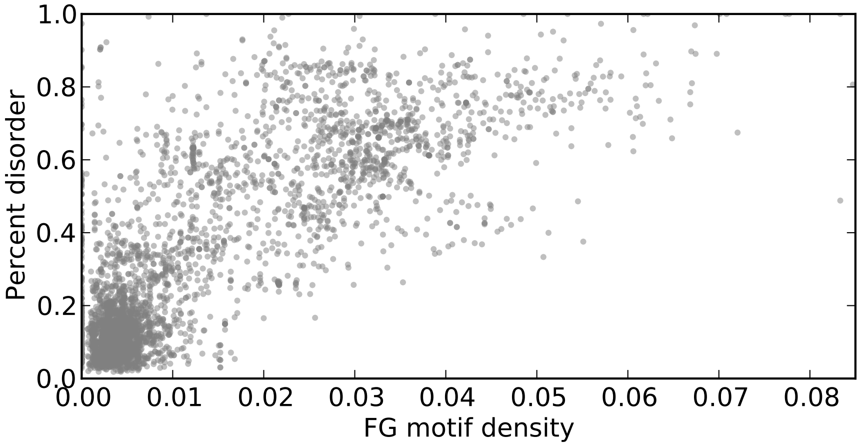

Figure 3.6: Continuous distribution of nups across NUP in terms of percent disorder (y-axis) and FG density in FG motifs/AA (x-axis).

In contrast, when the NUP proteins were restricted to those with 400 or more AAs and with fragment proteins removed, two distinct groups of nucleoporins were found to arise naturally, a group with low percentage disorder and low FG density, and another group with relatively high percentage disorder and relatively high FG density. Inspection of the names of the proteins showed that the low percentage disorder and low FG density group consists of karyopherin and structural nups, while the other group conyained known FG nups. An exhaustive and systematic labeling of all nucleoporins for Saccharomyces Cerevisiae yeast and humans confirmed that known karyopherin and structural nups formed one group, while the other group consisted of known FG nups. We took proteins in the NUP group with greater than 0.15 FG/AA linear motif density and greater than 30% disorder to be the FG nups which we analyzed in this study.

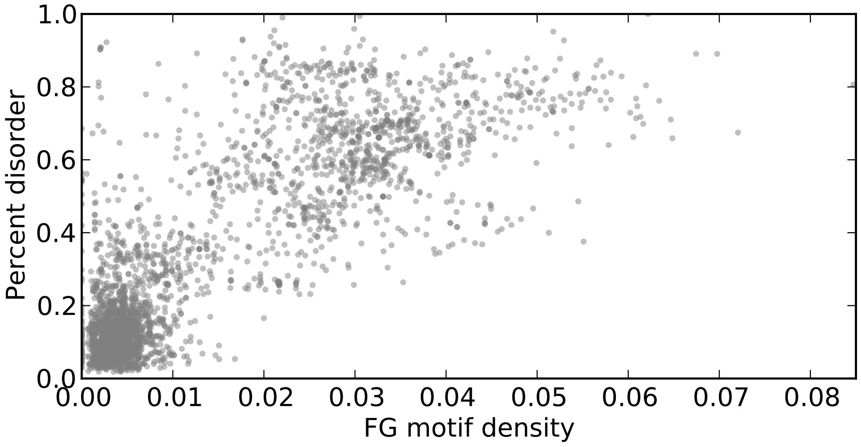

Figure 3.7: Natural split for nups for NUP restricted to proteins with greater than 400 AA at roughly greater than 10% FG/AA FG motif density and greater than 30% protein disorder.

Examining solely the Saccharomyces Cerevisiae nucleoporins confirms the split between FG nups and the structural/transport proteins as seen in Fig. S4.

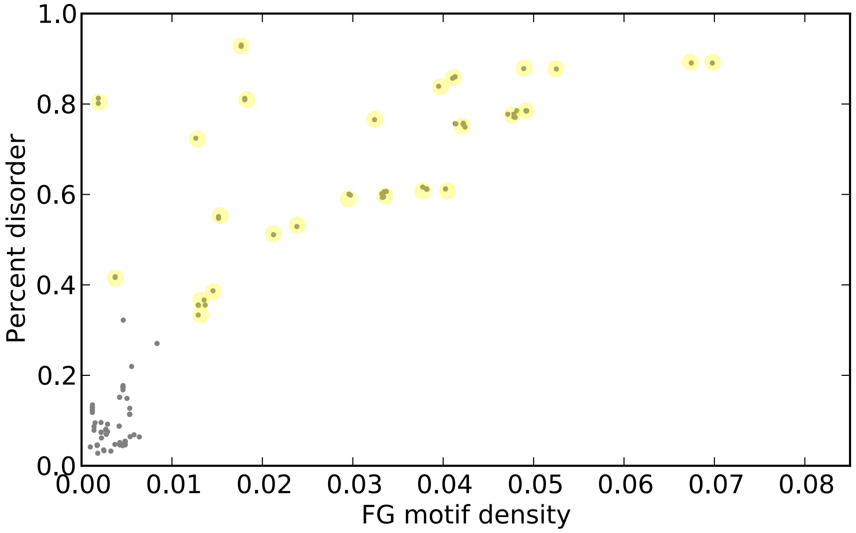

Figure 3.8: Natural split for Baker's Yeast, with 400 AA restriction. Yellow circles highlights refer to known FG nups while grey dots which are not highlighted represent known structural/transport proteins.

Examining solely the human nucleoporins confirms the split between FG nups and the structural/transport proteins as seen in Fig. S5.

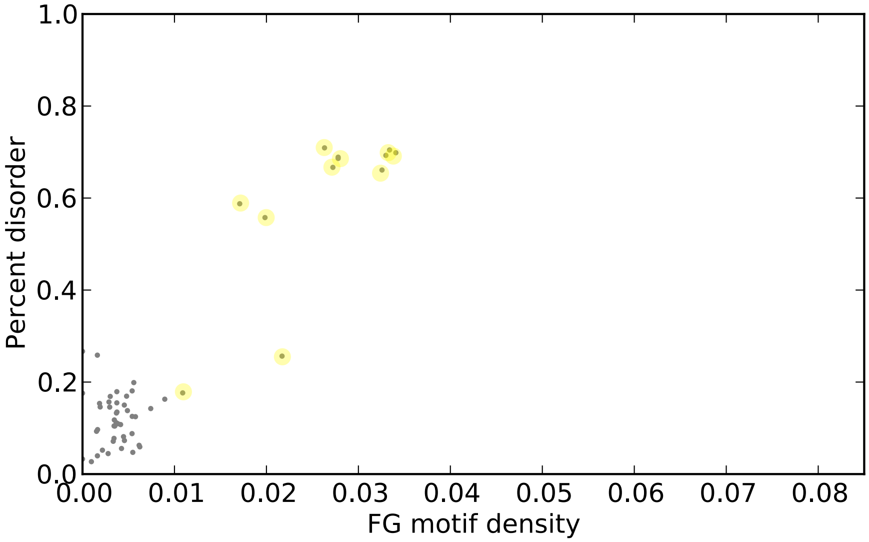

Figure 3.9: Natural split for humans, with 400 AA restriction. Yellow circles highlights refer to known FG nups while grey dots which are not highlighted represent known structural/transport proteins.

3.5.3 Disordered regions of FG nups which exclude FG clusters but contain charge clusters are neutral on average

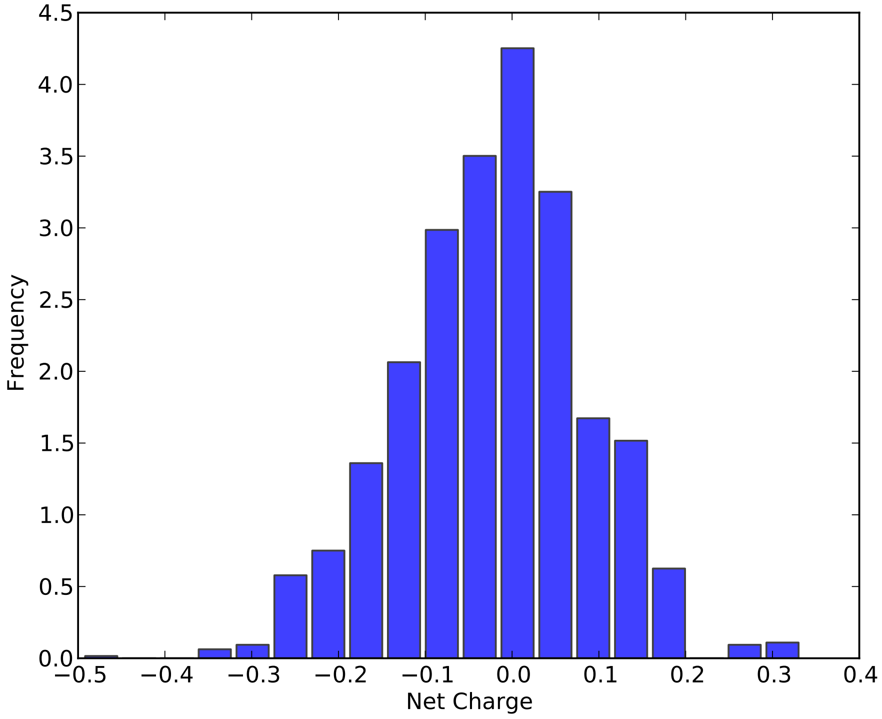

Figure 3.10: Net charge/AA for charged disordered charged regions of FG nups. The net charge for disordered charged regions of FG nups histogram is symmetric around zero net charge indicating that on average these regions are not selected for any net charge and on average are charge neutral.

3.5.4 Net charge of folded regions of FG nups

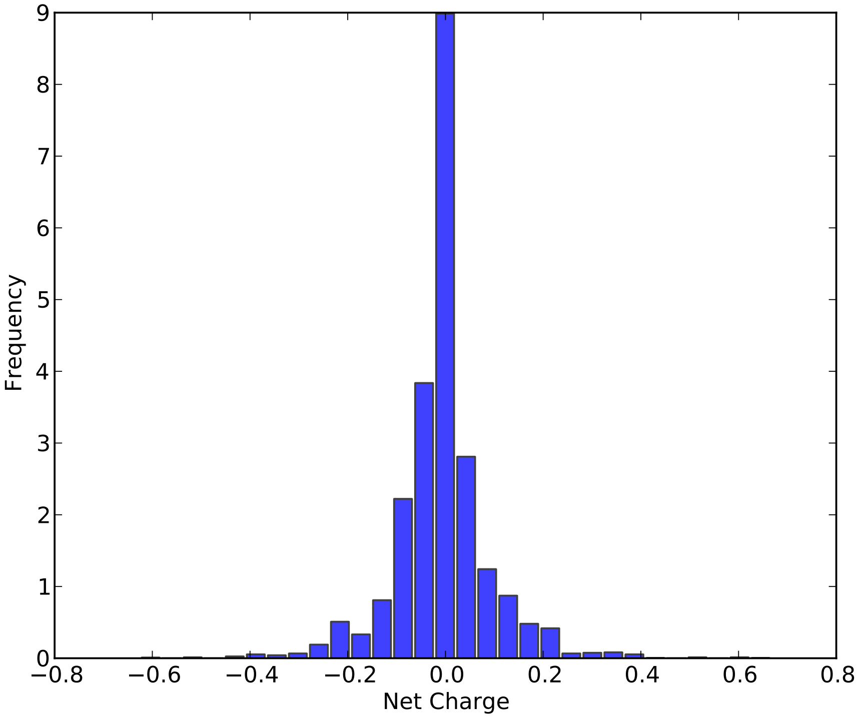

Figure 3.11: Net charge/AA for folded regions of FG nups is symmetric around zero net charge indicating that on average these regions are not selected for any net charge and on average are neutrally charged.

3.5.5 Clustering among individual FG nups

The 1,167 FG nups analyzed were clustered using the PreDeCon algorithm [51] over a five dimensional space consisting of FG-Charge cluster overlap in FG nups, FG-Charge cluster polarity, Folded-Disordered region polarity, percent of disordered region composed of charged AA clusters, and topological complexity. Setting the minimum cluster size to 50 proteins allowed for a large scale overview of how the 1167 FG nups organize themselves in this five dimensional space, which resulted in four major groupings of FG nups.

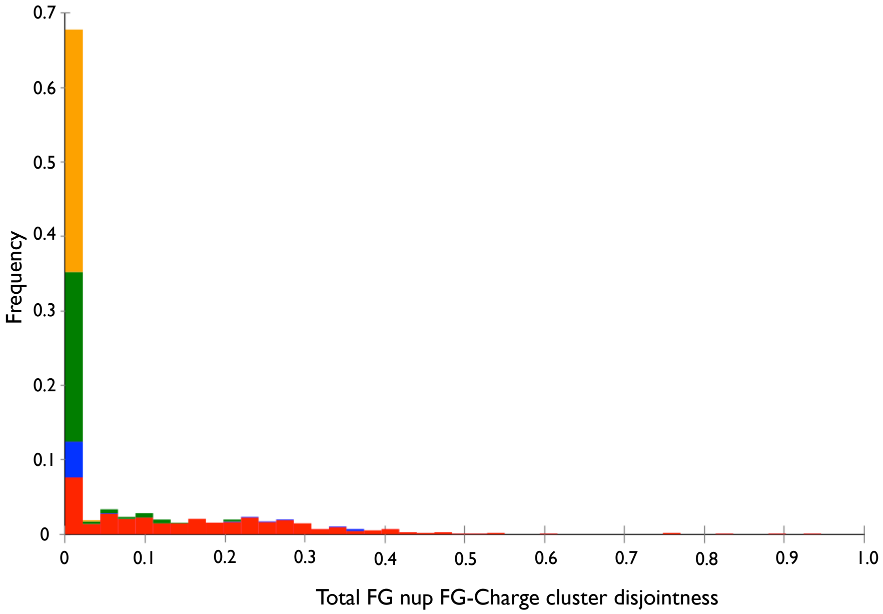

Figure 3.12: Histogram of FG-charge cluster percentage overlap in FG nups. Density clustering separates FG nups into 4 distinct groups, which are labeled, yellow, green, red and blue. Along the disjointness axis (x-axis) the red, green, and blue cluster groups aggregate at very high levels of disjointness, while the red group tends to have moderate levels of overlap.

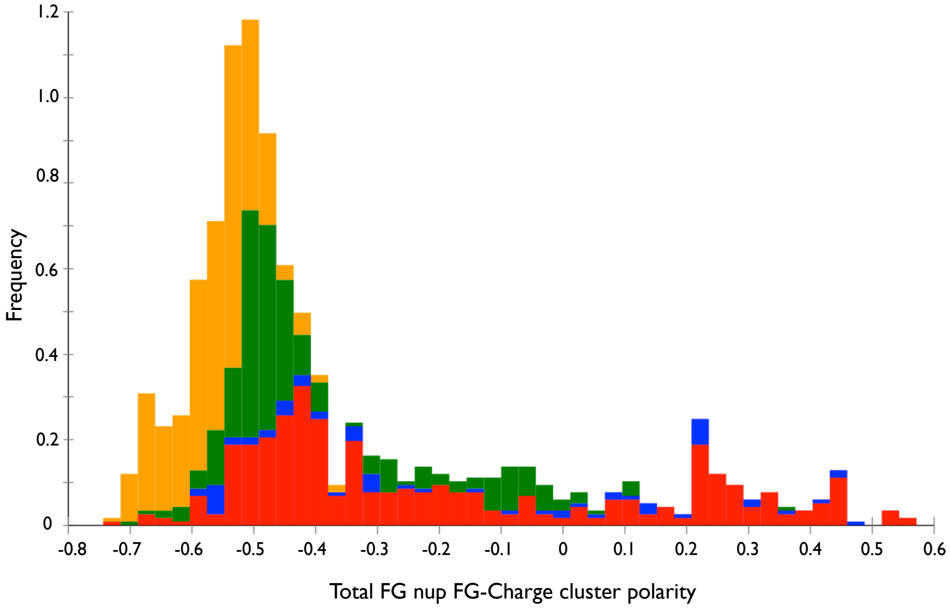

Figure 3.13: Histogram of FG-Charge cluster polarity. Density clustering separates FG nups into 4 distinct groups, which are shown in yellow, green, red and blue. Along the FG-Charge polarity axis positive polarity nups are nearly entirely from the red group, while other groups have significant negative polarity.

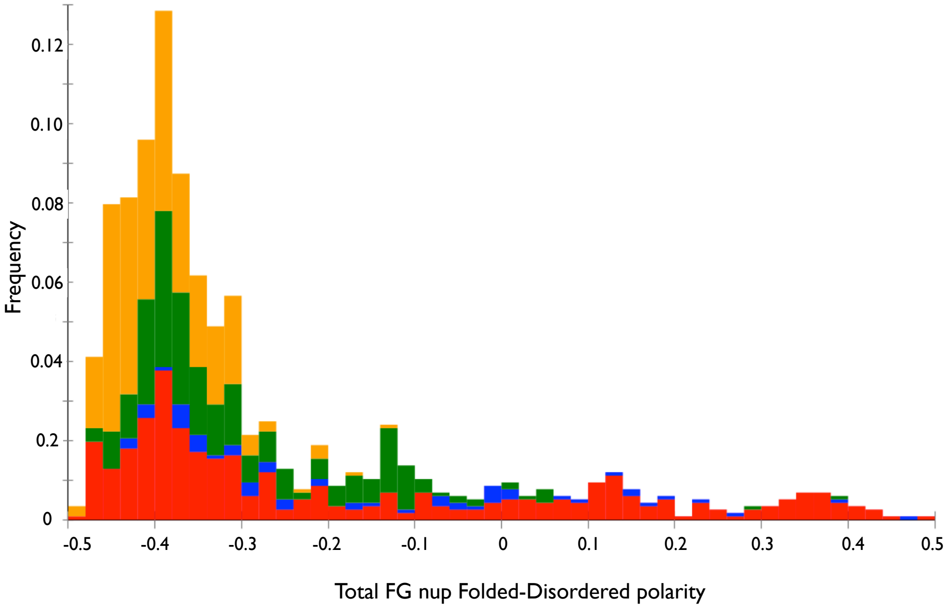

Figure 3.14: Histogram of FG nup Folded-Disordered region polarity. Density clustering separates FG nups into 4 distinct groups, which are labeled, yellow, green, red and blue. The positive polarity nups are nearly entirely from the red group, while other groups have significant negative polarity.

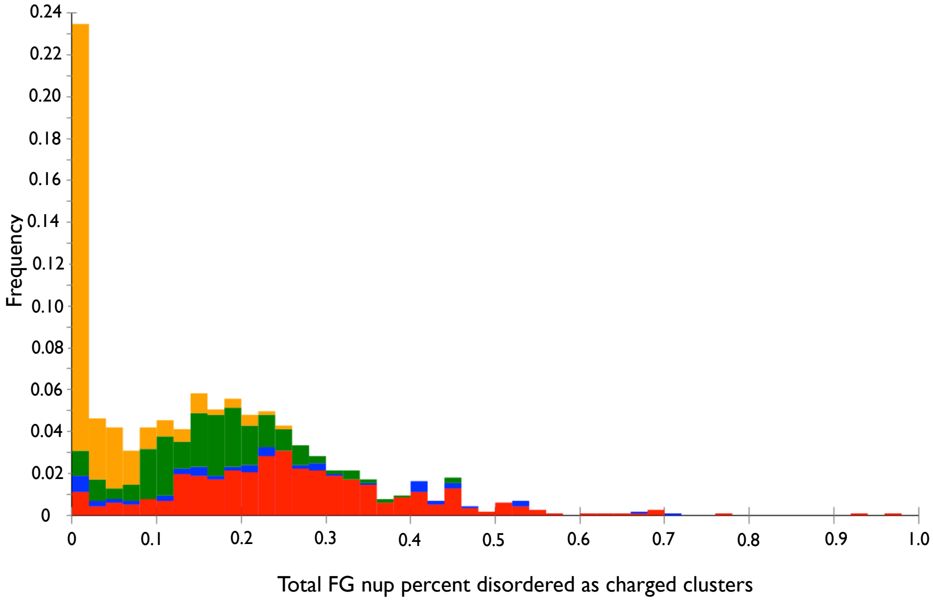

Figure 3.15: Histogram of the percent of disordered region composed of charged AA clusters for FG nups. Density clustering separates FG nups into 4 distinct groups, shown in yellow, green, red and blue. Along the percentage charged cluster axis the yellow group few charged amino acid clusters in the disordered regions, while the other three groups have significant charged disordered regions.

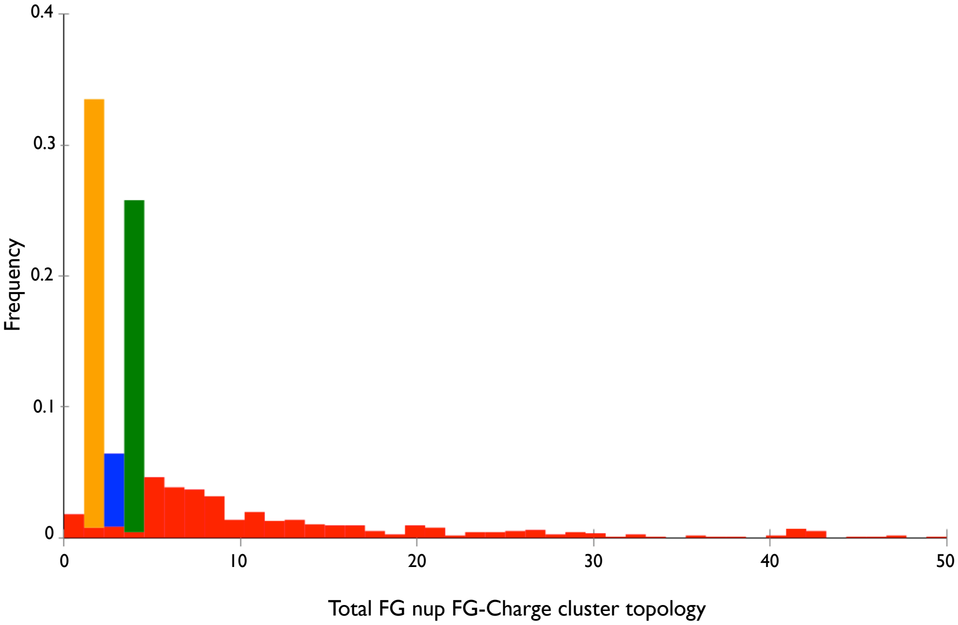

Figure 3.16: Histogram of the topological complexity of FG nups. Density clustering separates FG nups into 4 distinct groups, shown in yellow, green, red and blue. Along the topological complexity axis the yellow group has exclusively a topological complexity of 1, the blue group a topological complexity of 2, the green group a topological complexity of 3, while the red group has a topological complexity distributed over a range of higher values.

3.5.6 Overview of FG nups in each clustering group for Humans and S. cerevisiae with representative visualizations

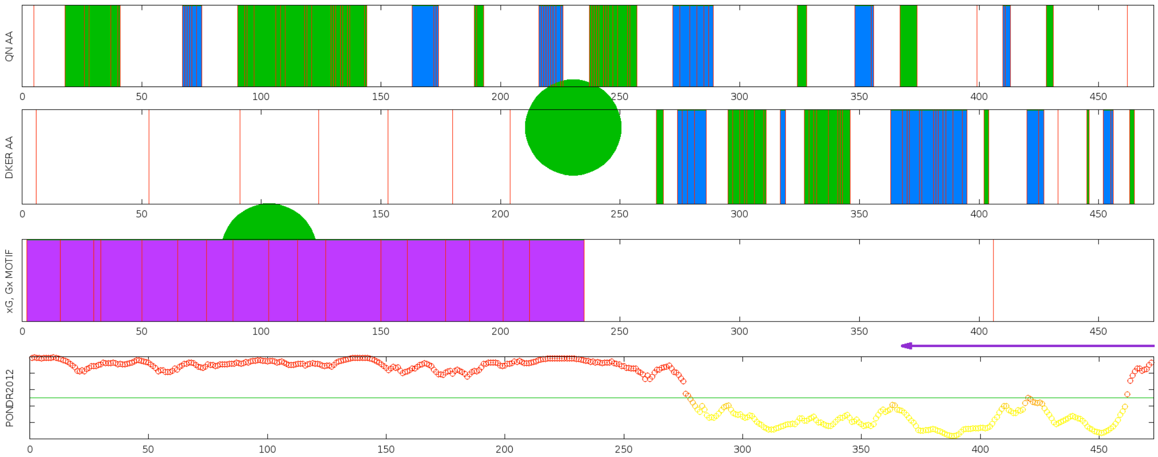

Figure 3.17: FG nucleoporin from S. cerevisiae, Nup49, from the yellow group. Amino acid sequence number is shown along the x-axis for all sub-charts. The first chart starting from the top shows QN amino acids as red vertical lines and their clusters colored alternately blue and green as a control. Similarly colored is the second chart which shows charged amino acids, while the third chart has FG motifs represented as red vertical lines with clusters represented by alternating purple and cyan regions. The fourth chart displays the propensity for protein disorder for a given AA as predicted by PONDR, with red representing high propensity and yellow representing low propensity. Green circles represent centers of masses of cluster regions and the purple arrow indicates disordered region to folded region polarity.

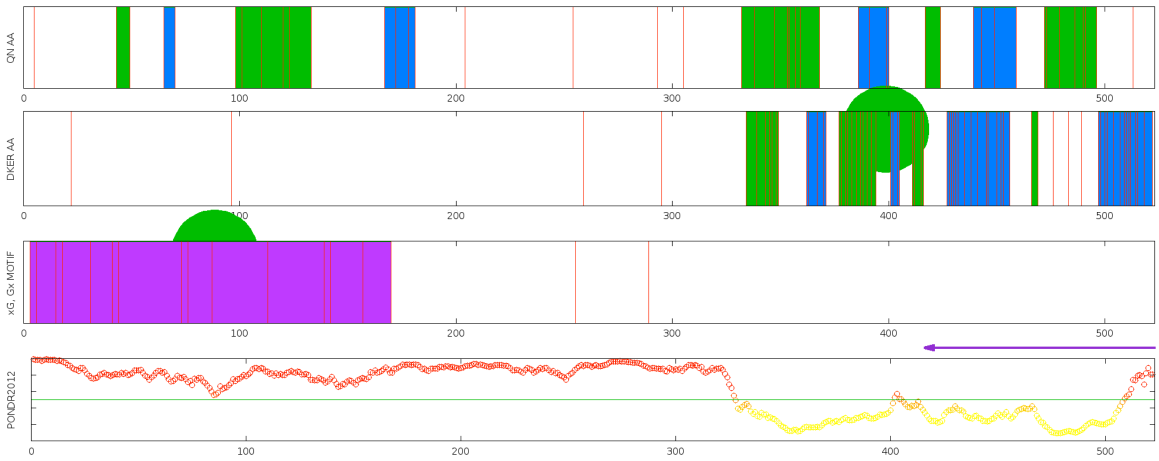

Figure 3.18: FG nucleoporin from Homo sapiens, Nup62, from the yellow group. Amino acid sequence number is shown along the x-axis for all sub-charts. The first chart starting from the top shows QN amino acids as red vertical lines and their clusters colored alternately blue and green as a control. Similarly colored is the second chart which shows charged amino acids, while the third chart has FG motifs represented as red vertical lines with clusters represented by alternating purple and cyan regions. The fourth chart displays the propensity for protein disorder for a given AA as predicted by PONDR, with red representing high propensity and yellow representing low propensity. Green circles represent centers of masses of cluster regions and the purple arrow indicates disordered region to folded region polarity.

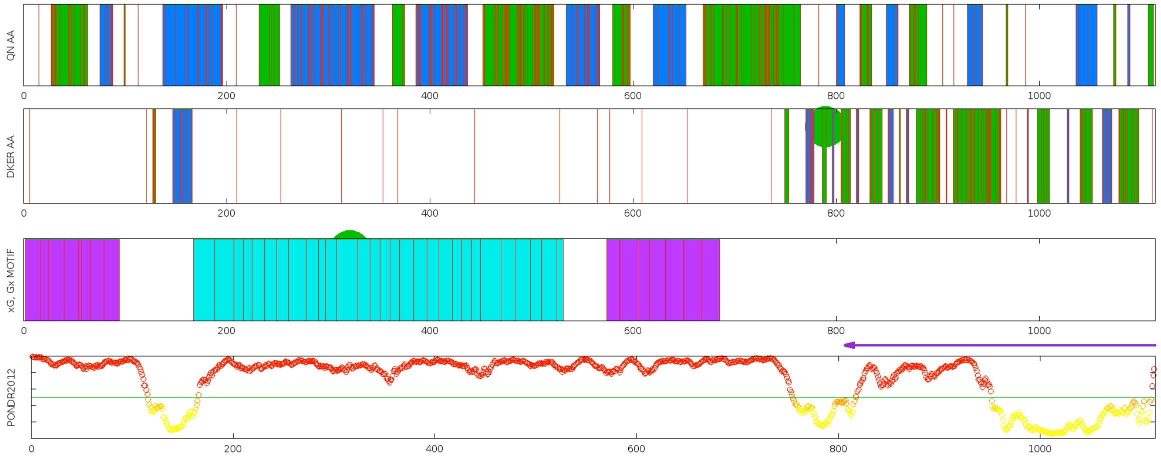

Figure 3.19: FG nucleoporin from S. cerevisiae, Nup116, from the green group. Amino acid sequence number is shown along the x-axis for all sub-charts. The first chart starting from the top shows QN amino acids as red vertical lines and their clusters colored alternately blue and green as a control. Similarly colored is the second chart which shows charged amino acids, while the third chart has FG motifs represented as red vertical lines with clusters represented by alternating purple and cyan regions. The fourth chart displays the propensity for protein disorder for a given AA as predicted by PONDR, with red representing high propensity and yellow representing low propensity. Green circles represent centers of masses of cluster regions and the purple arrow indicates disordered region to folded region polarity.

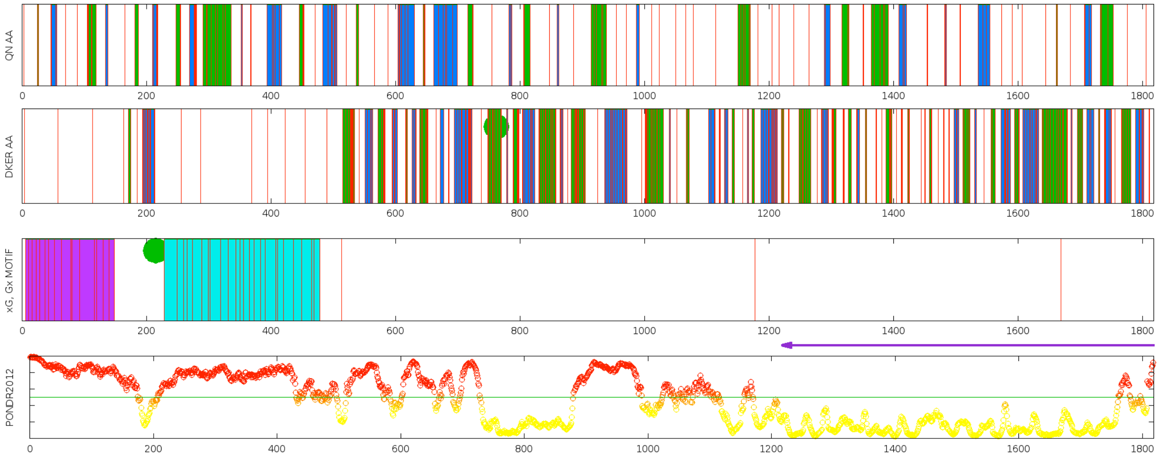

Figure 3.20: FG nucleoporin from Homo sapiens, Nup98, from the green group. Amino acid sequence number is shown along the x-axis for all sub-charts. The first chart starting from the top shows QN amino acids as red vertical lines and their clusters colored alternately blue and green as a control. Similarly colored is the second chart which shows charged amino acids, while the third chart has FG motifs represented as red vertical lines with clusters represented by alternating purple and cyan regions. The fourth chart displays the propensity for protein disorder for a given AA as predicted by PONDR, with red representing high propensity and yellow representing low propensity. Green circles represent centers of masses of cluster regions and the purple arrow indicates disordered region to folded region polarity.

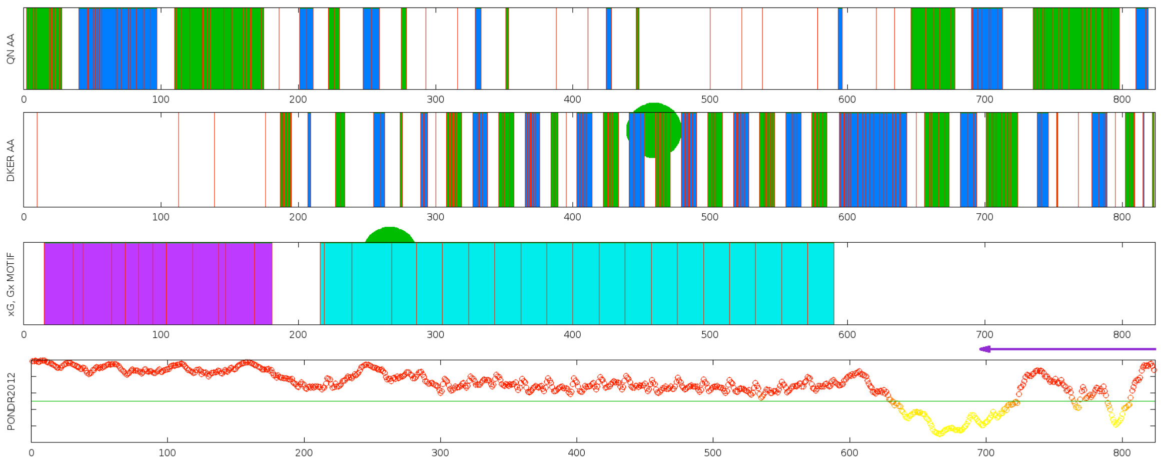

Figure 3.21: FG nucleoporin from S. cerevisiae, Nsp1, from the red group. Amino acid sequence number is shown along the x-axis for all sub-charts. The first chart starting from the top shows QN amino acids as red vertical lines and their clusters colored alternately blue and green as a control. Similarly colored is the second chart which shows charged amino acids, while the third chart has FG motifs represented as red vertical lines with clusters represented by alternating purple and cyan regions. The fourth chart displays the propensity for protein disorder for a given AA as predicted by PONDR, with red representing high propensity and yellow representing low propensity. Green circles represent centers of masses of cluster regions and the purple arrow indicates disordered region to folded region polarity.

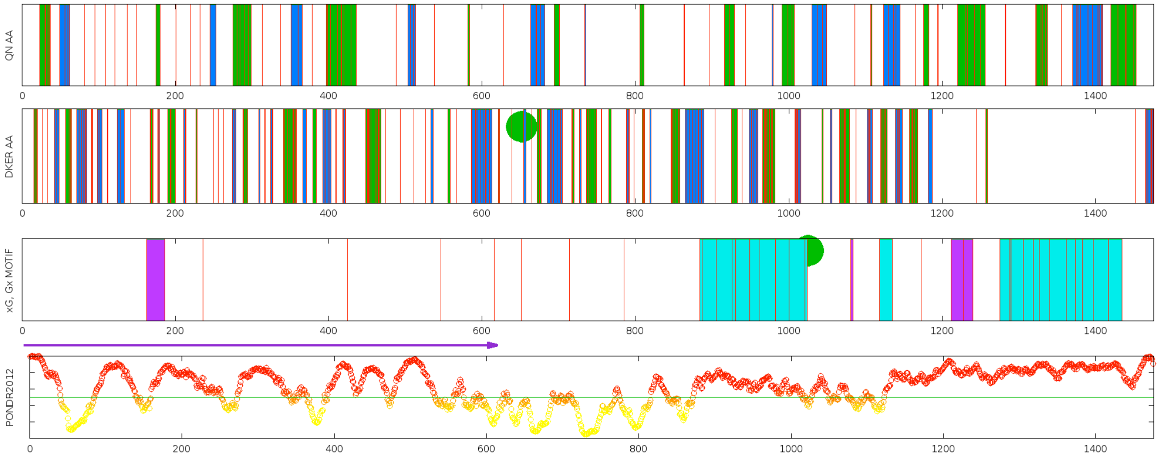

Figure 3.22: FG nucleoporin from Homo sapiens, Nup153, from the red group. Amino acid sequence number is shown along the x-axis for all sub-charts. The first chart starting from the top shows QN amino acids as red vertical lines and their clusters colored alternately blue and green as a control. Similarly colored is the second chart which shows charged amino acids, while the third chart has FG motifs represented as red vertical lines with clusters represented by alternating purple and cyan regions. The fourth chart displays the propensity for protein disorder for a given AA as predicted by PONDR, with red representing high propensity and yellow representing low propensity. Green circles represent centers of masses of cluster regions and the purple arrow indicates disordered region to folded region polarity.

Following is a list of well known FG nups from Humans and S. cerevisiae and their corresponding FG nup cluster.

List of proteins in the Red group:

tr—B3KUL1—B3KUL1 HUMAN cDNA FLJ40132 fis, clone TESTI2012155, highly similar to NUCLEOPORIN-LIKE PROTEIN RIP OS=Homo sapiens PE=2 SV=1

tr—B4DIK2—B4DIK2 HUMAN cDNA FLJ60565, highly similar to Nuclear pore complex protein Nup153 OS=Homo sapiens PE=2 SV=1

tr—B4DHC5—B4DHC5 HUMAN cDNA FLJ57927, highly similar to Nucleoporin-like protein RIP OS=Homo sapiens PE=2 SV=1

sp—P35658—NU214 HUMAN Nuclear pore complex protein Nup214 OS=Homo sapiens GN=NUP214 PE=1 SV=2

sp—P49790—NU153 HUMAN Nuclear pore complex protein Nup153 OS=Homo sapiens GN=NUP153 PE=1 SV=2

sp—P49792—RBP2 HUMAN E3 SUMO-protein ligase RanBP2 OS=Homo sapiens GN=RANBP2 PE=1 SV=2 sp—P52594—AGFG1 HUMAN Arf-GAP domain and FG repeat-containing protein 1 OS=Homo sapiens GN=AGFG1 PE=1 SV=2

sp—Q9UKX7—NUP50 HUMAN Nuclear pore complex protein Nup50 OS=Homo sapiens GN=NUP50 PE=1 SV=2

sp—Q96HA1—P121A HUMAN Nuclear envelope pore membrane protein POM 121 OS=Homo sapiens GN=POM121 PE=1 SV=2

tr—Q7Z743—Q7Z743 HUMAN Nucleoporin 153kDa OS=Homo sapiens GN=NUP153 PE=2 SV=1

tr—F6QR24—F6QR24 HUMAN Nuclear pore complex protein Nup153 OS=Homo sapiens GN=NUP153 PE=4 SV=1

tr—A6ZU88—A6ZU88 YEAS7 Nuclear pore complex subunit OS=Saccharomyces cerevisiae (strain YJM789) GN=NUP145 PE=4 SV=1

tr—A6ZPT6—A6ZPT6 YEAS7 Nuclear pore complex subunit OS=Saccharomyces cerevisiae (strain YJM789) GN=NSP1 PE=4 SV=1

tr—A7A1L6—A7A1L6 YEAS7 Nucleoporin OS=Saccharomyces cerevisiae (strain YJM789) GN=NUP2 PE=4 SV=1

tr—A6ZZP8—A6ZZP8 YEAS7 Nuclear pore complex subunit OS=Saccharomyces cerevisiae (strain YJM789) GN=NUP100 PE=4 SV=1

tr—A6ZVG2—A6ZVG2 YEAS7 Nucleoporin OS=Saccharomyces cerevisiae (strain YJM789) GN=NUP159 PE=4 SV=1

tr—B3LJE4—B3LJE4 YEAS1 Nucleoporin NUP1 OS=Saccharomyces cerevisiae (strain RM11-1a) GN=SCRG 01496 PE=4 SV=1

tr—B3LG91—B3LG91 YEAS1 Nucleoporin NUP42 OS=Saccharomyces cerevisiae (strain RM11-1a) GN=SCRG 00328 PE=4 SV=1

tr—B3LGY3—B3LGY3 YEAS1 Nucleoporin ASM4 OS=Saccharomyces cerevisiae (strain RM11-1a) GN=SCRG 00585 PE=4 SV=1

tr—B3LHF7—B3LHF7 YEAS1 Nucleoporin NUP145 OS=Saccharomyces cerevisiae (strain RM11-1a) GN=SCRG 01090 PE=4 SV=1

tr—B3RHK8—B3RHK8 YEAS1 Nucleoporin OS=Saccharomyces cerevisiae (strain RM11-1a) GN=SCRG 04280 PE=4 SV=1

tr—B3LTW2—B3LTW2 YEAS1 Nucleoporin OS=Saccharomyces cerevisiae (strain RM11-1a) GN=SCRG 05287 PE=4 SV=1

tr—B3LR23—B3LR23 YEAS1 Nucleoporin NUP100/NSP100 OS=Saccharomyces cerevisiae (strain RM11-1a) GN=SCRG 03952 PE=4 SV=1

tr—Q6FPU8—Q6FPU8 CANGA Similar to uniprot—P14907 Saccharomyces cerevisiae YJL041w NSP1 OS=Candida glabrata (strain ATCC 2001 / CBS 138 / JCM 3761 / NBRC 0622 / NRRL Y-65) GN=CAGL0J00781g PE=4 SV=1

sp—P14907—NSP1 YEAST Nucleoporin NSP1 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=NSP1 PE=1 SV=1

sp—P20676—NUP1 YEAST Nucleoporin NUP1 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=NUP1 PE=1 SV=1 sp—P32499—NUP2 YEAST Nucleoporin NUP2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=NUP2 PE=1 SV=2

sp—P40477—NU159 YEAST Nucleoporin NUP159 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=NUP159 PE=1 SV=1

sp—P49686—NUP42 YEAST Nucleoporin NUP42 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=NUP42 PE=1 SV=1

sp—P49687—NU145 YEAST Nucleoporin NUP145 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=NUP145 PE=1 SV=1

sp—Q05166—NUP59 YEAST Nucleoporin ASM4 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=ASM4 PE=1 SV=1

sp—Q03790—NUP53 YEAST Nucleoporin NUP53 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=NUP53 PE=1 SV=1

tr—Q6FPU8—Q6FPU8 CANGA Similar to uniprot—P14907 Saccharomyces cerevisiae YJL041w NSP1 OS=Candida glabrata (strain ATCC 2001 / CBS 138 / JCM 3761 / NBRC 0622 / NRRL Y-65) GN=CAGL0J00781g PE=4 SV=1

tr—C7GW26—C7GW26 YEAS2 Nup145p OS=Saccharomyces cerevisiae (strain JAY291) GN=NUP145 PE=4 SV=1

tr—C8ZBH3—C8ZBH3 YEAS8 Nsp1p OS=Saccharomyces cerevisiae (strain Lalvin EC1118 / Prise de mousse) GN=EC1118 1J11 2157g PE=4 SV=2

tr—C8Z8F6—C8Z8F6 YEAS8 Nup145p OS=Saccharomyces cerevisiae (strain Lalvin EC1118 / Prise de mousse) GN=EC1118 1G1 1959g PE=4 SV=1

tr—E7Q648—E7Q648 YEASB Nup100p OS=Saccharomyces cerevisiae (strain FostersB) GN=FOSTERSB 2851 PE=4 SV=1

tr—E7KNE8—E7KNE8 YEASL Nup145p OS=Saccharomyces cerevisiae (strain Lalvin QA23) GN=QA23 1700 PE=4 SV=1

tr—E7NJQ4—E7NJQ4 YEASO Nsp1p OS=Saccharomyces cerevisiae (strain FostersO) GN=FOSTERSO 2551 PE=4 SV=1

tr—E7KEX2—E7KEX2 YEASA Nup100p OS=Saccharomyces cerevisiae (strain AWRI796) GN=AWRI796 2893 PE=4 SV=1

tr—G2WGX7—G2WGX7 YEASK K7 Nsp1p OS=Saccharomyces cerevisiae (strain Kyokai no. 7 / NBRC 101557) GN=K7 NSP1 PE=4 SV=1

tr—G2WE08—G2WE08 YEASK K7 Nup145p OS=Saccharomyces cerevisiae (strain Kyokai no. 7 / NBRC 101557) GN=K7 NUP145 PE=4 SV=1

tr—H0GG71—H0GG71 9SACH Nup145p OS=Saccharomyces cerevisiae x Saccharomyces kudriavzevii VIN7 GN=VIN7 1730 PE=4 SV=1

tr—H0GJ66—H0GJ66 9SACH Nup100p OS=Saccharomyces cerevisiae x Saccharomyces kudriavzevii VIN7 GN=VIN7 2942 PE=4 SV=1

List of proteins in the Blue group:

sp—O15504—NUPL2 HUMAN Nucleoporin-like protein 2 OS=Homo sapiens GN=NUPL2 PE=1 SV=1 tr—Q3B7J4—Q3B7J4 HUMAN Nucleoporin like 2 OS=Homo sapiens GN=NUPL2 PE=2 SV=1

tr—B2R7I1—B2R7I1 HUMAN cDNA, FLJ93452, highly similar to Homo sapiens nucleoporin like 2 (NUPL2), mRNA OS=Homo sapiens PE=2 SV=1

tr—B4DWF8—B4DWF8 HUMAN cDNA FLJ53414, highly similar to Nuclear pore complex protein Nup98-Nup96precursor OS=Homo sapiens PE=2 SV=1

List of proteins in the Yellow group:

sp—P37198—NUP62 HUMAN Nuclear pore glycoprotein p62 OS=Homo sapiens GN=NUP62 PE=1 SV=3

tr—B4E0K0—B4E0K0 HUMAN cDNA FLJ52820, highly similar to Nucleoporin p58/p45 OS=Homo sapiens PE=2 SV=1

tr—C9JDH3—C9JDH3 HUMAN Nucleoporin p58/p45 OS=Homo sapiens GN=NUPL1 PE=4 SV=2

tr—E7Q443—E7Q443 YEASB Nup57p OS=Saccharomyces cerevisiae (strain FostersB) GN=FOSTERSB 1849 PE=4 SV=1

sp—P48837—NUP57 YEAST Nucleoporin NUP57 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=NUP57 PE=1 SV=1

tr—A6ZUD2—A6ZUD2 YEAS7 Nucleoporin OS=Saccharomyces cerevisiae (strain YJM789) GN=NUP57 PE=4 SV=1

tr—B3LIB0—B3LIB0 YEAS1 Nucleoporin OS=Saccharomyces cerevisiae (strain RM11-1a) GN=SCRG 00900 PE=4 SV=1

tr—B5VJ65—B5VJ65 YEAS6 YGR119Cp-like protein OS=Saccharomyces cerevisiae (strain AWRI1631) GN=AWRI1631 73440 PE=4 SV=1

tr—G2WEK1—G2WEK1 YEASK K7 Nup57p OS=Saccharomyces cerevisiae (strain Kyokai no. 7 / NBRC 101557) GN=K7 NUP57 PE=4 SV=1

tr—C7GUH3—C7GUH3 YEAS2 Nup57p OS=Saccharomyces cerevisiae (strain JAY291) GN=NUP57 PE=4 SV=1

tr—E7KCJ0—E7KCJ0 YEASA Nup57p OS=Saccharomyces cerevisiae (strain AWRI796) GN=AWRI796 1881 PE=4 SV=1

tr—C8Z910—C8Z910 YEAS8 Nup57p OS=Saccharomyces cerevisiae (strain Lalvin EC1118 / Prise de mousse) GN=EC1118 1G1 4335g PE=4 SV=1

tr—H0GGN3—H0GGN3 9SACH Nup57p OS=Saccharomyces cerevisiae x Saccharomyces kudriavzevii VIN7 GN=VIN7 1905 PE=4 SV=1

tr—E7NHZ5—E7NHZ5 YEASO Nup57p OS=Saccharomyces cerevisiae (strain FostersO) GN=FOSTERSO 1834 PE=4 SV=1

tr—H0GUL6—H0GUL6 9SACH Nup49p OS=Saccharomyces cerevisiae x Saccharomyces kudriavzevii VIN7 GN=VIN7 7019 PE=4 SV=1

tr—E7KCC2—E7KCC2 YEASA Nup49p OS=Saccharomyces cerevisiae (strain AWRI796) GN=AWRI796 1636 PE=4 SV=1 tr—E7QEJ1—E7QEJ1 YEASZ Nup49p OS=Saccharomyces cerevisiae (strain Zymaflore VL3) GN=VL3 1625 PE=4 SV=1

tr—A6ZU14—A6ZU14 YEAS7 Nuclear pore complex subunit OS=Saccharomyces cerevisiae (strain YJM789) GN=NUP49 PE=4 SV=1

tr—B3LHM4—B3LHM4 YEAS1 Nucleoporin NUP49/NSP49 OS=Saccharomyces cerevisiae (strain RM11-1a) GN=SCRG 01163 PE=4 SV=1

tr—C7GPN5—C7GPN5 YEAS2 Nup49p OS=Saccharomyces cerevisiae (strain JAY291) GN=NUP49 PE=4 SV=1

tr—C8Z879—C8Z879 YEAS8 Nup49p OS=Saccharomyces cerevisiae (strain Lalvin EC1118 / Prise de mousse) GN=EC1118 1G1 1068g PE=4 SV=1

tr—E7KN89—E7KN89 YEASL Nup49p OS=Saccharomyces cerevisiae (strain Lalvin QA23) GN=QA23 1628 PE=4 SV=1

tr—E7LUN9—E7LUN9 YEASV Nup49p OS=Saccharomyces cerevisiae (strain VIN 13) GN=VIN13 1619 PE=4 SV=1

tr—H0GG42—H0GG42 9SACH Nup49p OS=Saccharomyces cerevisiae x Saccharomyces kudriavzevii VIN7 GN=VIN7 1656 PE=4 SV=1

sp—Q02199—NUP49 YEAST Nucleoporin NUP49/NSP49 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=NUP49 PE=1 SV=1

tr—G2WDT3—G2WDT3 YEASK K7 Nup49p OS=Saccharomyces cerevisiae (strain Kyokai no. 7 / NBRC 101557) GN=K7 NUP49 PE=4 SV=1

tr—B5VM90—B5VM90 YEAS6 YKL068Wp-like protein OS=Saccharomyces cerevisiae (strain AWRI1631) GN=AWRI1631 111600 PE=4 SV=1

tr—B5VPH5—B5VPH5 YEAS6 YMR047Cp-like protein OS=Saccharomyces cerevisiae (strain AWRI1631) GN=AWRI1631 131890 PE=4 SV=1

tr—G2WHW4—G2WHW4 YEASK K7 Nup100p OS=Saccharomyces cerevisiae (strain Kyokai no. 7 / NBRC 101557) GN=K7 NUP100 PE=4 SV=1

3.5.7 List of the 252 FG nup species analyzed

Acromyrmex echinatior, Aedes aegypti, Ailuropoda melanoleuca, Ajellomyces capsulata, Ajellomyces dermatitidis, Albugo laibachii, Amblyomma maculatum, Anguilla japonica, Anopheles darlingi, Anopheles gambiae, Arabidopsis lyrata subsp. lyrata, Arabidopsis thaliana, Arthrobotrys oligospora, Arthroderma benhamiae, Arthroderma gypseum, Arthroderma otae, Ashbya gossypii, Aspergillus clavatus, Aspergillus flavus, Aspergillus kawachii, Aspergillus niger, Aspergillus oryzae, Aspergillus terreus, Aureococcus anophagefferens, Babesia bovis, Batrachochytrium dendrobatidis, Blastocystis hominis, Bos taurus, Botryotinia fuckeliana, Branchiostoma floridae, Brugia malayi, Caenorhabditis brenneri, Caenorhabditis briggsae, Caenorhabditis elegans, Caenorhabditis japonica, Caenorhabditis remanei, Callithrix jacchus, Camponotus floridanus, Candida albicans, Candida dubliniensis, Candida glabrata, Candida parapsilosis, Candida tropicalis, Canis familiaris, Cavia porcellus, Chaetomium globosum, Chaetomium thermophilum, Chaetomium thermophilum var. thermophilum, Chlorella variabilis, Ciona intestinalis, Ciona savignyi, Clavispora lusitaniae, Clonorchis sinensis, Coccidioides posadasii, Colletotrichum graminicola, Colletotrichum higginsianum, Coprinopsis cinerea, Cordyceps militaris, Cricetulus griseus, Cryptococcus gattii serotype B, Cryptococcus neoformans var. neoformans serotype D, Cryptosporidium hominis, Cryptosporidium muris, Cryptosporidium parvum, Culex quinquefasciatus, Danaus plexippus, Danio rerio, Daphnia pulex, Daucus carota, Debaryomyces hansenii, Dicentrarchus labrax, Dictyostelium discoideum, Dictyostelium fasciculatum, Dictyostelium purpureum, Drosophila ananassae, Drosophila erecta, Drosophila grimshawi, Drosophila melanogaster, Drosophila mojavensis, Drosophila persimilis, Drosophila pseudoobscura pseudoobscura, Drosophila sechellia, Drosophila simulans, Drosophila virilis, Drosophila willistoni, Drosophila yakuba, Ectocarpus siliculosus, Emericella nidulans, Encephalitozoon cuniculi, Encephalitozoon intestinalis, Entamoeba dispar, Entamoeba histolytica, Equus caballus, Eremothecium cymbalariae, Exophiala dermatitidis, Fusarium oxysporum, Gallus gallus, Gasterosteus aculeatus, Giardia intestinalis, Glarea lozoyensis, Glossina morsitans morsitans, Gorilla gorilla gorilla, Grosmannia clavigera, Harpegnathos saltator, Heterocephalus glaber, Homo sapiens, Hordeum vulgare var. distichum, Hydra vulgaris, Hypocrea atroviridis, Hypocrea jecorina, Hypocrea virens, Ichthyophthirius multifiliis, Kazachstania africana, Kluyveromyces lactis, Laccaria bicolor, Lachancea thermotolerans, Latimeria chalumnae, Leishmania braziliensis, Leishmania donovani, Leishmania infantum, Leishmania major, Leishmania mexicana, Leptosphaeria maculans, Loa loa, Lodderomyces elongisporus, Loxodonta africana, Macaca fascicularis, Macaca mulatta, Magnaporthe oryzae, Malassezia globosa, Melampsora larici-populina, Metarhizium acridum, Metarhizium robertsii, Meyerozyma guilliermondii, Micromonas sp., Mixia osmundae, Monodelphis domestica, Monosiga brevicollis, Mus musculus, Mycosphaerella graminicola, Myotis lucifugus, Naegleria gruberi, Naumovozyma castellii, Naumovozyma dairenensis, Nectria haematococca, Nematostella vectensis, Neosartorya fischeri, Neosartorya fumigata, Neospora caninum, Neurospora crassa, Neurospora tetrasperma, Nomascus leucogenys, Nosema ceranae, Oikopleura dioica, Oncorhynchus mykiss, Ornithorhynchus anatinus, Oryctolagus cuniculus, Oryza sativa subsp. indica, Oryza sativa subsp. japonica, Oryzias latipes, Ostreococcus lucimarinus, Otolemur garnettii, Pan troglodytes, Paracoccidioides brasiliensis, Paramecium tetraurelia, Pediculus humanus subsp. corporis, Penicillium chrysogenum, Penicillium marneffei, Perkinsus marinus, Phaeodactylum tricornutum, Phaeosphaeria nodorum, Physcomitrella patens subsp. patens, Phytophthora infestans, Phytophthora ramorum, Phytophthora sojae, Picea sitchensis, Pichia angusta, Pichia pastoris, Pichia sorbitophila, Piriformospora indica, Plasmodium falciparum, Plasmodium knowlesi, Plasmodium vivax, Plasmodium yoelii yoelii, Podospora anserina, Polysphondylium pallidum, Pongo abelii, Populus trichocarpa, Postia placenta, Pristionchus pacificus, Puccinia graminis f. sp. tritici, Pyrenophora teres f. teres, Pyrenophora tritici-repentis, Rattus norvegicus, Rhodotorula glutinis, Ricinus communis, Saccharomyces cerevisiae, Salmo salar, Salpingoeca sp., Sarcophilus harrisii, Scheffersomyces stipitis, Schistosoma mansoni, Schizophyllum commune, Schizosaccharomyces japonicus, Schizosaccharomyces pombe, Sclerotinia sclerotiorum, Serpula lacrymans var. lacrymans, Sordaria macrospora, Sorghum bicolor, Spathaspora passalidarum, Sporisorium reilianum, Strongylocentrotus purpuratus, Sus scrofa, Taeniopygia guttata, Takifugu rubripes, Talaromyces stipitatus, Tetrahymena thermophila, Tetraodon nigroviridis, Tetrapisispora phaffii, Thalassiosira pseudonana, Theileria annulata, Theileria parva, Thielavia heterothallica, Thielavia terrestris, Torulaspora delbrueckii, Toxoplasma gondii, Tribolium castaneum, Trichinella spiralis, Trichomonas vaginalis, Trichophyton equinum, Trichophyton rubrum, Trichophyton tonsurans, Trichophyton verrucosum, Trichoplax adhaerens, Trypanosoma brucei brucei, Trypanosoma brucei gambiense, Trypanosoma congolense, Trypanosoma cruzi, Trypanosoma vivax, Tuber melanosporum, Uncinocarpus reesii, Ustilago maydis, Vanderwaltozyma polyspora, Verticillium albo-atrum, Verticillium dahliae, Vitis vinifera, Volvox carteri, Xenopus laevis, Xenopus tropicalis, Yarrowia lipolytica, Zea mays, and Zygosaccharomyces rouxii.

3.5.8 Spatial Sequence Correlation of FG and GF motifs

Within disordered regions of FG nups we found that the FG and GF motifs are colocalized. Excluding the motifs that were isolated from other motifs by greater than 100 AAs as noise, we found the distance to the nearest motif neighbor of a specific type from a given motif of a specific type. This measurement, averaged over all motifs in all the FG nups studied in this paper, resulted in the following average nearest neighbor separation values of 18.3 AAs between FG motifs, 24.0 AAs between GF motifs, and 21.7 AA between GF motifs and FG motifs.

Chapter 4

MOLECULAR MODELING OF FG NUCLEOPORINS

4.1 Coarse Grained Simulations

While all atom Molecular Dynamics (MD) models can accurately reproduce the local secondary structure and nature of tertiary contacts for proteins, it is not computationally feasible to study even a single FG nup's properties due to their large size (∼1000 AAs) and the long timescales involved (μs). Characterizing the overall structure formed by a hundred odd FG nups becomes next to impossible. To overcome these issues we use a Coarse Grained (CG) model [27] which is detailed enough so that it is able to reasonably reproduce secondary and tertiary structure and even fold small proteins *ab initio* starting from extended states, but simple enough such that multi-protein systems over microsecond timescales are computationally feasible. This CG model attempts to reproduce the energy landscape of real proteins by representing each amino acid by multiple CG interaction sites and by using specially developed interactions between these sites. The physical potentials between these CG interaction centers is derived from measuring the forces present in highly accurate all atom simulations of diverse protein sequences, following the multiscale coarse-graining (MS-CG) method [75]. This method makes no assumptions about the type of interactions between CG sites, and additionally implicitly includes multibody correlations [76] in the resulting implicit solvent model with an effective CG potential. In this CG model, secondary and tertiary structure emerge naturally through the comprehensiveness of the physical model, and it is general enough to simulate proteins of arbitrary amino acid sequence.

Typically, CG potentials for a specific protein simulation are scaled such that model results match well with experiment [27,77]. This was achieved by multiplying the entire CG potential by a constant scaling factor. Experimentally measured properties of disordered FG nups have yet to converge, therefore we developed a number of CG models with different scaling factors (α, β, γ), with each scaling factor corresponding to a different class of experimental results. FRET measurements of human nup153 by Milles *et al* [78] have shown the disordered regions of this protein to consist of nearly entirely compact collapsed coil configurations. The CG model α was therefore developed to mimic these results, scaled such that it produces a collapsed coil (molten globule) [79] for the disordered region of nup153. (Table 4.1 compares model and experimental radii in the Supplementary Material) Other experiments have shown nup153 to take on more extended configurations, forming relatively extended brush configurations which undergo reversible collapse in the presence of transport factors [80]. CG model γ was therefore developed to mimic these results, scaled such that it produces a relaxed coil [79] for the disordered region of nup153, Table 4.1. Bead-halo experiments by Yamada *et al* [3] indicate nups are biphasic, with stalk domains non-interacting in general while on the other hand FG domains are "sticky" to other FG domains. FG nups were also found to be biphasic in terms of polymer structure, with stalk domains having expanded polymer conformations, while FG domains had more compact polymer conformations. A scaling factor β precisely in the middle of the α and γ scalings (Methods) was found to produce such biphasicness, with this scaled CG model labeled β.

The FG nups we simulated were taken from evolutionary distant *S. cerevisiae* (Nsp1, Nup1, Nup100, and Nup116) and *H. sapiens* (Nup98 and Nup153), focusing on long FG nups with large disordered regions which have a "biphasic" sequence structure which can be separated into an "FG domain" and "stalk domain". An FG domain corresponds to a sequence region with a low concentration of charged amino acids and a high concentration of FG motifs, while the stalk domain corresponds to a disordered sequence region with a high concentration of charged amino acids (nup specific amino acid domain definitions and disordered domains used are defined in Supplementary Material and Yamada *et al* [3]).

4.1.1 Structure and dynamics of single nups

Polypeptide chains representing the disordered regions of various FG nups from *S. cerevisiae* and humans were started from an initially fully extended configuration and equilibrated before the start of simulation. Fig. 4.1 shows a typical snapshot of one such nup, Nsp1, after collapse and equilibration. The snapshot seems to indicate that these proteins can exhibit distinctly different kinds of polymeric structure along a single chain with regions that look like swollen chains to other parts that resemble collapsed coils. In general, the proteins seemed to adopt a block polymer structure to varying degrees, dependent on the form of CG model used. Amino acid regions of the FG nups simulated and domain definitions used are in Supplementary Material, Table 4.2.

To quantitatively characterize FG nup structure, we define the contact probability between a pair of amino acids as the percentage of time the corresponding CG backbone beads are within 1.6 nm of each other. The diblock polymer structure of these FG nups is particularly clear from the contact probability maps showing the contact probability over all possible amino acid contact pairs (Fig. 4.2). In many cases these show a universal block diagonal structure with a part that is almost diagonal - representing an extended stalk and a square block that is uniformly lit up indicating a collapsed disordered structure - the tip. The lack of contacts between the blocks indicates the blocks maintain their distinct identities over time. The diblock structure was most strongly seen in model β. Model α often produced significant contacts between FG and stalk domains which destroys diblock structure as the two domains often interact, while model γ produced nups with little difference between FG and stalk domains with the entire FG nups representing unstructured extended homogenous polymers. Under all model assumptions, individual FG nups were highly dynamic and disordered. FG domains across all simulated nups maintained a relatively smooth block contact structure, with contacts between amino acids rapidly exchanging and no folded constant contacts visible. The consistency in the diblock structure that is exhibited across the different nups indicates that there must be some underlying general features in the sequence structure that is maintained. In particular, this would suggest that disruption or randomization of the sequence would result in a loss of such structure. To test this we repeated the simulations with a shuffled version of the amino acid sequence of Nsp1 which resulted in the loss of diblock structure (Supplementary Material, Fig. 4.6). This clearly suggests that assuming nup sequences have an effectively homogeneous character could be problematic and that the biphasic sequence structure is responsible for producing two distinct types of dynamic domains via weak but specific interactions among portions of FG nups. The sequence features of the FG and stalk domains which delineate these domains is discussed in the Bioinformatics section later in this paper and in Ando *et al* [81].

An earlier hypothesis that FG nups can form a "tree" like structure was based on the experimentally measured properties of *fragments* of FG nups from *S. cerevisiae* [3]. Here, our simulation results on full length FG nups support this hypothesis by demonstrating that it is clearly possible that individual fragment properties combine to form true diblock copolymers. Additionally, the CG model used here is the first such model capable of reproducing the observed interactions [3] between different fragments from different regions of the FG nups.

To quantitatively characterize the levels of disorder within distinct regions of the contact probability maps we derived looping probabilities from the maps for different domains of the FG nups. We define the looping probability $P_l(s)$ to be the probability that two monomers of size b, separated by a contour length of sb along the amino acid chain are within a spatial distance $r < R_{cut}$. In our analysis $R_{cut} = 1.6$ nm. For two monomers of an *ideal chain*, the looping probability for a given monomer separation distance sb, is given by the integral of the probability distribution of end-to-end distances up to R_{cut}

$$P_l(s) = \int_{r<R_{cut}} \mathrm{d}V \left(\frac{3}{2\pi sb^2}\right)^p \exp\left(\frac{-3r^2}{2sb^2}\right) \tag{4.1}$$

with exponent $p = 3/2$ [82]. An extended chain would have a power of $p = 9/5$ and a collapsed coil would have a $p = 1$, with appropriately scaled constant pre-factors.

Additionally we quantitatively determined the statistical significance of the block structure of contact maps in the γ model, as they appear to disappear in this model. At a 40 amino acid separation distance we measured the average contact probability and standard deviation between all possible amino acid combinations for the shuffled Nsp1 sequence that is shown in the Supplementary Material. Measuring the same 40 amino acid separation contact probabilities for the γ model contact maps we determined that the average of this contact probability over all contact maps in this model was only 1.02 standard deviations away from the corresponding value for the unstructured shuffled Nsp1 sequence. In contrast, the same contact probability averaged over the β model contact maps was 3.58 standard deviations away from the shuffled Nsp1 sequence value.

Fig. 4.1 shows a log-log plot of the average measured looping probability as a function of s for the all FG and stalk domains simulated, with a numerically fitted theoretical looping probability (for the power p) over the amino acids separations greater than 5 and less than 65 monomers. Qualitatively, a high looping exponent power p for a particular domain would imply that that part of the polymer is relatively extended spatially, as the contact probability falls off more rapidly as s is increased within that domain.

As can be seen in Fig. 4.1, FG domains have looping exponents significantly lower than those of stalk domains, indicating the constantly large difference in ensemble structure between these two types domains. For the α, β, and γ CG models the average fitted scaling exponents for the FG domains were 1.16, 1.38, 1.81 respectively. For the stalk domains the average fitted scaling exponents were 2.08, 2.58, and 2.82 for the different models respectively. We also note an important caveat that these simulations of individual FG nups by themselves were done ignoring the effects of the crowded pore environment [13] and other FG nups. Although we analyze aggregates of interacting FG nups in the next section, our simulations do not include the crowding effects from transport factors, cargo, and other molecules in the cellular milieu which can significantly modulate FG nup behavior [83].

4.1.2 Structure and dynamics of nup rings

To look at the structures formed by multiple interacting nups *in vivo*, we performed simulations of rings of Nsp1 and Nup100 (identical sequences as in the individual nup simulations) grafted in an eightfold symmetric manner to the inner surface of a neutral, non-reactive cylinder, with the grafted end locations chosen in a manner consistent with the grafting orientation and FG nup anchor domain positions within the nuclear pore. For a single ring of FG nup "trees" we chose the dimensions of the enclosing cylindrical boundary to be (5 nm axial length, 50 nm diameter) representative of the geometry of the NPC as shown in Fig. 4.9 [3].

Fig. 4.3 shows the total mass density of the simulated rings of Nsp1 FG nups. The mass density clearly reveals a dense central plug connected by peripheral cables to the pore walls in the α and β models, with a plug mass density which is in all scenarios much smaller than the mass density of folded proteins of $\sim$1400 mg/cm^3 [84]. The mass density can be tied directly to the microphase separation of the different blocks of the natively unfolded proteins. One block phase (the FG domains) separated along the center of the channel while the other block (the stalk regions) aggregated along the periphery (Fig. 4.3). This microphase separation could explain the observation of a plug like structure in cryo-electron microscopic examination of NPCs [85]. This plug like structure is currently not recognized as a structural element of the nuclear pore complex and is rather rationalized as an artifact of cargo in transit [86]. Interestingly, in the γ model which represents an extended coil scenario there is no microphase separation present, reminiscent of the virtual gate model.

Spatial net charge density of the ring of Nsp1 nups mirrors closely the mass density of the FG domain, as FG domains are in general positively charged [81]. This produces a striking effect in the α and β models, with a strongly positively charged plug region forming which has a net charge density much higher than considered earlier by Ribbeck *et al* [87] and Szleifera *et al* [26], with the same possible functionality but much greater attraction to negatively charged transport factors due to radial concentration towards the center of the NPC channel.

In contrast to nups like Nsp1 and Nup1 which have a long stalk region and comparatively small sticky tip, other FG nups, like Nup100 have large sticky tips and comparatively short stalk regions. We simulated rings of Nup100 in an identical manner to Nsp1, resulting in similar results except for a key difference that Nup100 rings in isolation are unable to close in the α and β models, unlike Nsp1 which can close the pore in these scenarios (Fig. 4.10). In the next section, we use polymer brush modeling to understand the structure of Nup100 rings in the presence of closed Nsp1 rings, and find a cooperative effect where Nup100 rings will close in the presence of closed Nsp1 rings.

We did not observe any rigid contacts from the ring simulations which lasted for a time in excess one microsecond, so we rule out the existence of amyloid or "rigid" hydrogel structures which have contacts on these timescales in the β model used. This can also be seen by the high similarity of the contact map for an individual free Nsp1 (Fig. 4.2) and the contact map for a grafted Nsp1 (Fig. 4.7) in the ring conformation. Even in the compact α model, the dense FG domain "central plug" contains individual FG domains which maintain their disordered state and rapidly exchange contacts to form block diagonal contact maps for the full Nsp1 nup.

4.2 Polymer brush morphology and dynamics

To understand the underlying physics that governs the formation of different types of mesoscale structures and possible transitions between them, we turn to polymer physics modeling. Here we introduce a simple polymer brush model that treats an idealized polymer brush constructed of diblock polymers representing the FG nups. We assume that the nups can be described by neutral excluded volume polymers (representing the stalk domains) with an appropriate length and stiffness, and grafted to the inside of a cylinder of diameter equal to that of the nuclear pore. We then incorporate the biphasic structure of the nups in the α and β models by modeling the ends of the polymers as structureless cohesive blobs that correspond to the cohesive FG domains at the nup tips. The resulting configuration is a polymer brush as shown schematically in Fig. 4.4A.

The question now becomes - what is the height H of this brush? Do the nups extend all the way into the center of the pore and fill the space or are they collapsed along the sidewalls or do they form a brush with an intermediate height that still leaves an open transport conduit along the middle of the pore? The equilibrium height of this brush can be derived from a minimization of the free energy of the brush which has contributions coming from (i) entropic stretching of the stalks (ii) the excluded volume interactions (steric hindrance) between the stalks and (iii) the cohesive energy interactions between the tip blobs. Even within this simple model the curvature of the grafting surface and the cohesive tips result in the brush height being given by the solution to a non-linear differential equation. We can however make another mean-field approximation to arrive at an algebraic form for the effective free energy per chain (Supplementary Material);

$$\frac{F}{k_BT} \sim (H/a)^{5/2}N^{-3/2} + \frac{a^{5/2}N^{3/2}RH^{-1/2}}{d^2(2R-H)} - \frac{\epsilon\delta^2 R}{d^2(R-H)} \tag{4.2}$$

where ϵk_BT is the effective cohesive energy between a pair of blobs. The first two terms of the free energy follow directly from prior work on polymer brushes in cylinders [88, 89].

First we specifically choose one of the Nups to model : Nsp1. The extended stalk domains are modeled by excluded volume polymer chains with sizes as measured from the β simulation, which determines the Kuhn length a and effective polymerization number N (Supplementary Material). The cohesive domains at the tip are expected to be in a small, compact, collapsed state. We model these as featureless sticky blobs of size (diameter) $\delta = 3.6$ nm, twice the radius of gyration as measured in the β-CG model simulations. We assume that the average spacing between grafting points of Nsp1 along the pore wall is $d \sim 10$ nm, which we estimate by assuming that there are 32 copies [5] of Nsp1 anchored along the inner wall in a symmetric fashion along the geometry used in Yamada et al [3] (Fig. S4). The pore radius is taken to be $R \sim 25$ nm.

Fig. 4.4 B shows a plot of the free energy as a function of the brush height for different values of the cohesive energy with all the other parameters being set by the values for Nsp1 that were detailed above. The minima in these curves correspond to equilibrium values of the brush height. For no cohesion ($\epsilon = 0$) we see that the equilibrium actually corresponds to a brush height that is about half the pore radius, implying that if cohesion between the tips is lost there will be a wide-open channel down the center of the pore. We call this the "open" state of the gate. As the cohesive energy is increased, we see a new minimum developing at $H \sim R - \delta$, which is the state where the stalks are extended so that the cohesive blobs are at the center, thus completely closing the conduit. This we will refer to as the "closed" state. This indicates that there will be a transition between the closed and open states as the cohesion energy is varied. Thus the biphasic functional arrangement of domains with an extended stalk and cohesive tip leads to a brush structure that can be switched between two states of the brush: open and closed, as shown in Fig. 4.4. With reference to gated transport, it is known that transport factors, such as karyopherin/importin, bind to cohesive FG motifs [90, 91]. This provides a mechanism to disrupt the inter-molecular cohesiveness between nup tips by the competitive binding of karyopherin. We propose that the binding of certain particular transport factors *in vivo* results in such a switch from the closed to open state forming the basis for our model. In other words specific types of transport mediated cargo complexes could have the ability to dissolve into the central plug that keeps the gate closed, thus effectively opening it up to the desired diameter. These types of cargo can then undergo one dimensional diffusion down the center of the channel. Enzymes on the nuclear side can cleave the transport factors, reducing the propensity of cargo to remain in the channel leading to escape to the nearest side - the nucleus. This picture works in the other direction with export signals as well and forms the basis of our *Diblock Copolymer Brush Gate* model. It is important to note that small cargo can also have alternate routes through the sides and this aspect will be fully explored in future work.

We also used polymer brush modeling to understand the structure of Nup100 rings in the presence of closed Nsp1 rings. The free energy per chain is the same as in Eq. 4.2, but with an additional Nup100-Nsp1 cohesive cross interaction term. Polymer sizes used were as measured in the Nup100 β-CG model simulations (Supplementary Material). The per chain cross interaction free energy term between the two different types of FG nups can be simply modeled as the overlap fraction of an individual Nup100 sticky tip to the Nsp1 sticky tips at the center of the pore, multiplied by the blob interaction energy of the Nup100 sticky tips, $\epsilon k_B T$. The total Nup100 free energy is shown in Fig. 4.11 for different values of the blob-blob interaction energy, with $\epsilon k_B T$ on the scale of the estimated self interaction energy of Nup100 sticky tips (Supplementary Material). Rings of Nup100 are unable to close the pore, except in the presence of an adjacent closed ring of Nsp1 nups. Given how our brush modeling and CG simulation show that rings of Nup100 are unlikely to close the nuclear pore except in the presence of other closed rings of FG nups, cooperativity between FG nups could play an important role in nucleocytoplasmic gating. This idealized analysis where only Nup100 and Nsp1 are present, ignores the excluded volume that 'shrub' FG nups like Nup49 and Nup57 along the pore wall could have in crowding out stalk domains and similar effects induced by cargo and transport factors, although we tentatively conclude that Nsp1 or a FG nup with similar properties could be principal to the closing or 'sealing' mechanism of the pore. Consistent with this hypothesis combinatorial deletions of FG nups of the type performed by Strawn et al [92], have shown perturbations in transport in mutants when Nsp1 or nups like Nsp1 are functionally impaired. Our idealized analysis also ignores possible compactification, extension, crowding, and cross-linking of FG nups via interactions with cargo and their transport factors, although the existence and/or nature of these effects *in vivo* remain unclear.

4.3 Bioinformatics

Our modeling connects the biphasic sequence structure across FG nups in *S. cerevisiae* to their diblock polymer properties which in turn results in a bistable polymer brush morphology. If this unique brush structure is critical for NPC gating it would imply that the biphasic sequence structure should be conserved across eukarya. To examine this assertion we analyzed FG motif density versus charged AA (amino acid) density across the disordered regions of the FG nup with the most FG repeats across ten diverse eukaryotic species, Fig. 4.5. We found that these nups all showed a common feature where, within each individual FG nup there are two domains, one high in FG motifs with another high in charged AAs, with both domains disjoint from one another along the amino acid sequence. One domain, labeled the "FG domain" [3], is low in charged AAs and high in FG motif repeats, while the other domain, the "stalk domain" [3], has a high charged AA density and typically a low FG repeat density with an anchor domain which attaches the FG nup to the inner pore surface. A quantitative bioinformatic analysis of the distribution of motifs along nup sequences by Ando *et al* [81] revealed a broad conservation across species of these bimodal sequence structures in FG nups. Our modeling work here suggests a possible explanation for the functional advantage gained by utilizing the bimodal sequence of FG nups in NPC transport regulation.

4.4 Discussion

Many experimental results regarding individual FG nup properties such as structure, location and cohesive properties of different domains of nups are in conflict. What happens when these FG nups are put together within the confines of the nuclear pore channel is even more uncertain. Here we attempted to understand how the overall architecture of the assembly of disordered FG nups in the NPC channel is governed by the properties of individual FG-nups that make up the assembly using large scale physically accurate coarse grained simulations. We calibrated our CG model against three main classes of experimental data on individual FG nup properties. We showed that, under certain conditions, a dynamic hydrogel-like structure does form, but only locally across spatial domains that have a high concentration of "FG domains". We do not observe any scenario where FG nups form a dense layer located along the pore wall with a relatively open center as seen in some simulations [24–26], highlighting the importance of protein model structure on results. Using a calibrated model where the FG nups are relatively compact, our results indicate that a Forest type model structure [3] emerges, with the "sticky tips" of FG nups coalescing in the center to form a hydrophobic plug flanked by an extended coil brush zone.

Some recent experimental studies have measured that certain forms of facilitated transport take place near the walls of the NPC. [93–97]. Facilitated transport through peripheral routes does not appear to be a natural outcome from previous simulation efforts of the NPC [23–26]. Such transport, however, appears to be supported by the α and β models where there exists a dense plug like structure along the center of the channel which can block some forms of transport, forcing smaller actively transported cargoes to take the peripheral route along the channel walls. Other cargoes, larger in size and which have yet to be experimentally studied, may contain the proper transport factors which can open up the central plug as our polymer modeling suggests.

Our polymer brush modeling indicates that for the critical "tree" like nups to functionally open and close the pore in the Forest model, their sequence must follow a specific arrangement. The motifs and charges should form a diblock pattern and have a fixed polarity with the extended stalk region coming after the anchor region followed by the cohesive FG regions. Our analysis of ten selected nups (with the largest number of FG motifs) across different species as well as earlier bioinformatic analysis of thousands of FG nups [81] clearly show that a biphasic sequence pattern is repeated across widely varying species, indicating that a Diblock Copolymer Brush Gating mode of transport may be a generic feature of certain types of nuclear import and export. Our modeling, therefore, sheds light on the underlying physics and biological function responsible for the presence of this biphasic pattern. Since the models α, β, γ can also be viewed as roughly simulating FG nup properties under differing solvent conditions, other models of transport which do not rely upon individual FG nup biphasic sequence structure may also emerge as limiting cases of our model in different effective solvent conditions. We imagine that the γ model loosely corresponds to poor solvent conditions because hydrophobic-hydrophobic interactions are weakened in poor solvent conditions, model γ represents a weakening of hydrophobic interactions, and that the FG nups are dominated by hydrophobic amino acids due to their compact size relative to neutral polymers of the same length. In the γ model the FG nups form a homogenous brush within the pore similar to the virtual gate polymer brush model, a gating mechanism the NPC may share with the Diblock Copolymer Brush Gate model under different types of solvent conditions.

Throughout this paper we have focused on the larger ”tree” type FG nups which have been shown to be critical for many forms of transport [92]. Some details of NPC structure such as cytoplasmic filaments, nuclear basket structure, and short ”shrub” FG nups which presumably lie along the pore wall have not been considered in our analysis. Since the cytoplasmic filaments and the basket are spatially well separated from the interior of the pore we anticipate that they would not interfere with the copolymer brush structure. Given that the properties of the shrub FG nups are similar to the sticky tip domains of tree FG nups in terms of bimodal adhesion [3], we anticipate that the shrub-sticky attractive tip interactions could help stabilize the wide open configurations of the copolymer brush in the presence of very large cargo in the middle of the pore. This sort of cooperativity could be relevant for the export of large ribonucleoprotein particles whose transport is facilitated by surface bound transport factors. In the work of Strawn et al [92] shrub like nups Nup49 and Nup57 were required for the viability of yeast cells, which indicates that the effects of shrubs are important and possibly necessary for the transport of large cargoes. It is interesting to note that we have shown that another form of cooperativity exists between different types of FG nups. Our modeling shows that the closure of the Nup100 ring requires the presence of a closed ring of Nsp1 or Nsp1 like nups. These and possibly other forms of cooperativity could play an important role in heightening the selectivity over a wider range of cargo sizes. Therefore, as is the case with numerous other biomolecular assemblies [98], cooperativity may be significant for NPC function.

Indirect evidence for the existence of a *Diblock Copolymer Brush Gate* mechanism *in vivo* include the fact that in mammals and yeast, weak alcohols such as hexanediol can ”loosen” the permeability barrier of the NPC at low concentrations (5%) [99, 100]. Also, exposed electrical charges in cargos bound to karyopherins can slow down their passage across the NPC [101]. Both these seem to indicate a hydrophobic plug but are also consistent with the hydrogel model. One way to truly differentiate between these models is to shorten/extend the stalk regions or to rearrange the cohesive regions within the same nups. We propose a plausible experiment where genetically engineered organisms have the anchor domains of their core FG nups moved away from their stalk domains and towards the end of their FG domains, a genetic rearrangement which we would predict to be lethal, while most other models, which assume homogenous FG nups, would predict identical function and viable organisms. Similarly, we also propose that genetically engineered organisms which have had their disordered FG nup sequence regions randomly shuffled should be lethal, while again most other models would predict identical function. Physically, we predict lethality in these genetic engineering experiments as they would destroy the specific microphase separation of the cylindrical brush we predict would be key to the transport of certain types of important cargoes.

Our results not only increase the number of physical mechanisms by which we can understand nucleocytoplasmic transport, but also allow us to extend our model to designing and optimizing novel forms of biomimetic transport. Biomimetic membranes [102] have a wide variety of applications ranging from chemical and biological separation and purification [103], as a platform for analytical detection of antioxidants, antibiotics, antiviral, psychotropic and other substances, drug delivery [104] as well as self-contained reactors and mock cells [105, 106]. The regulated cross-membrane transport of material in such systems is of vital importance and our investigations of a well-defined bio-compatible mechanism should be very useful. In general, any diblock polymers which follow the sequence determined polymer properties of FG nups (excluded volume polymer 'stalk', with collapsed sticky 'tips') could be used to design biomimetic pores which regulate traffic using our Diblock Copolymer Brush Gate mechanism. In addition, amino acid sequences of disordered proteins can be modified to change their overall polymer properties to be similar to FG nups in our diblock copolymer pore model, allowing similar regulatory pores to be created in many different contexts. Indeed, even though the NPC transport is poorly understood, there have already been efforts to make biomimetic gates inspired by the NPC [107, 108].

4.5 Methods

The CG model of Hills *et al* [27], was run using the large-scale atomic/molecular massively parallel simulator (LAMMPS) [109] software suite. Using the three separate calibrated CG models with α, β, γ equal to 3.35, 4.30, and 5.26 respectively at 300 K, simulated polypeptide chains representing the disordered regions of various FG nups from *S. cerevisiae* and humans were equilibrated for 1 microsecond, using starting conditions which consisted of fully extended protein configurations for the simulations without the cylindrical boundary conditions. Disordered regions of FG nups which were simulated used the nup specific definitions defined in Yamada *et al* [3] and the Supplementary Material. For the pore-like simulations with cylindrical boundary conditions, initial nup configurations (identical sequences as in the individual nup simulations) were as depicted in the Supplementary Material. During the 1 microsecond equilibration, all individual FG nups collapsed from their initially fully extended configuration in under 300 ns. After equilibration data was then taken from a 4 microsecond production run. Post simulation analysis was done to determine the average radius of gyration of different domains as defined in [3]. Contact proximity between two given side-chain beads was calculated by counting the number of times side-chains were within 16 Angstroms of each other in the trajectory snapshots which were saved every 100 picoseconds during the 4 microseconds of production simulation, then normalizing by the number of snapshots. Initial starting conditions for the ring simulations can be see in Supplementary Material. Heat maps of the density of FG motifs and charged amino acids in the FG nup with the most FG repeats across ten different species we only applied to sequence regions which were predicted to be disordered via a PONDR-FIT score greater than 0.5 [110] and to have greater than 40 AAs. The resulting binary densities of FG motif/AA and charged amino acid/AA were smoothed by a running average over a 20 AA window.

4.6 Supplementary Material

4.6.1 Cylindrical Brush Free Energy

Characterization of the NPC cylindrical polymer brush is achieved by deriving its total free energy per chain:

$$F_T = F_s + F_{ex} + F_{coh} \tag{4.3}$$

The first term in the RHS of this equation, F_s, is the stretching free energy which we model as originating from stalk part of the NPC brush only. The second term F_{ex} is the excluded volume free energy of this cylindrical brush, and the last term is the cohesive energy F_{coh} of the sticky tip FG domains.

4.6.2 Free energy of stalk chain stretching

The per chain stretching free energy is [111]:

$$F_s = k_bT(\#blobs_{stalk}) = k_bT \int_{stalk} \mathrm{d}n_b \tag{4.4}$$

With n_b the number of blobs per nup and where the stretching free energy per nup is equal to Boltzmann's constant times the temperature times the number of blobs in the stalk region of a nup.

The number of blobs in the stalk brush if we have a *flat brush* is:

$$n_b = L/\xi \tag{4.5}$$

With L the end to end distance of stalk chains and ξ the radius of the stalk brush blobs.

The number of monomers per extended chain is:

$$N = n_b(\xi/a)^{5/3} \tag{4.6}$$

With a equal to the length of monomers in the chains.

$$\Rightarrow N = (L/\xi)(\xi/a)^{5/3} \tag{4.7}$$

$$\Rightarrow \xi = (N/L)^{3/2}a^{5/2} \tag{4.8}$$

Now we can consider the cylindrical brush case. First we consider blob size changes along the stalk brush as a function of the contour length s, which is a function of monomer number m. i.e. m is the mth monomer starting from the direction of the center of the cylinder with $m = 0$ the free end of the stalk chain: In direct analogue to the flat brush case, Eq. 4.8:

$$\Rightarrow \xi(s) = (\frac{\mathrm{d}m}{\mathrm{d}s})^{3/2}a^{5/2} \tag{4.9}$$

We can now calculate the change in the number of blobs as a function of s, analogous to Eq. 4.5, as:

$$\Rightarrow \mathrm{d}n_b = \frac{\mathrm{d}s}{\xi(s)} \tag{4.10}$$

Which after substitution from Eq. 4.20 becomes [88]:

$$\Rightarrow \mathrm{d}n_b = \mathrm{d}s(\frac{\mathrm{d}s}{\mathrm{d}m})^{3/2}a^{-5/2} \tag{4.11}$$

We can now solve for the stretching free energy in Eq. 4.4

$$F_s = k_bT\int_{stalk}\mathrm{d}n_b = N_ck_bT\int\mathrm{d}s(\frac{\mathrm{d}s}{\mathrm{d}m})^{3/2}a^{-5/2} \tag{4.12}$$

$$= k_bT\int\mathrm{d}m(\frac{\mathrm{d}s}{\mathrm{d}m})(\frac{\mathrm{d}s}{\mathrm{d}m})^{3/2}a^{-5/2} \tag{4.13}$$

$$= k_bT\int\mathrm{d}m(\frac{\mathrm{d}s}{\mathrm{d}m})^{5/2}a^{-5/2} \tag{4.14}$$

Which after the mean field approximation of:

$$\frac{\mathrm{d}s}{\mathrm{d}m} = \frac{H}{N} \tag{4.15}$$

With H the height of the cylindrical stalk brush. Eq. 4.14 reduces to:

$$F_s = k_bTN(\frac{H}{N})^{5/2}a^{-5/2} \tag{4.16}$$

Where we can now conclude that the stretching free energy per chain is:

$$F_s = k_bT(\frac{(H/a)^{5/2}}{N^{3/2}}) \tag{4.17}$$

4.6.3 Free energy of excluded volume interactions

For a flat brush the free energy per chain of excluded volume interactions is (1):

$$F_e = \frac{1}{2}k_bT(\#blobs_{stalk})\rho_{blob} \tag{4.18}$$

With ρ_{blob} the volume fraction of blobs in the brush.

This can be generalized to a per chain free energy valid for cylindrical brushes:

$$F_s = \frac{1}{2}k_bT\int_{stalk}\mathrm{d}n_b\rho_{blob} \tag{4.19}$$

With n_b the number of blobs per nup. The local volume faction can be defined as:

$$\rho_{blob} = \frac{\mathrm{d}V_{blobs}}{\mathrm{d}V_{brush}} = \frac{\xi^3 N_c \mathrm{d}n_b}{2\pi s\mathrm{d}sL} \tag{4.20}$$

Which implies that:

$$F_s = \frac{1}{2}k_bT\int_{stalk}\mathrm{d}n_b\frac{\xi^3 N_c\mathrm{d}n_b}{2\pi s\mathrm{d}sL} \tag{4.21}$$

Which simplifies to:

$$F_s = \frac{1}{2}k_bT\int_{stalk}\mathrm{d}s\frac{\mathrm{d}n_b^2}{\mathrm{d}s^2}\frac{\xi^3 N_c}{2\pi sL} \tag{4.22}$$

Which after substitution for $\frac{\mathrm{d}n_b}{\mathrm{d}s}$ by Eq. 4.10 equals:

$$F_s = \frac{1}{2}k_bT\int_{stalk}\mathrm{d}s\frac{\xi N_c}{2\pi sL} \tag{4.23}$$

Which after substitution for ξ from Eq. 4.20:

$$F_s = \frac{1}{2}k_bT\int_{stalk}\mathrm{d}s\frac{(\frac{\mathrm{d}m}{\mathrm{d}s})^{3/2}a^{5/2}N_c}{2\pi sL} \tag{4.24}$$

Which equals:

$$F_s = \frac{1}{2}k_bT\int_{stalk}\mathrm{d}m\frac{\mathrm{d}s}{\mathrm{d}m}\frac{(\frac{\mathrm{d}s}{\mathrm{d}m})^{-3/2}a^{5/2}N_c}{2\pi sL} = k_bT\int_{stalk}\mathrm{d}m\frac{(\frac{\mathrm{d}s}{\mathrm{d}m})^{-1/2}a^{5/2}N_c}{\pi sL} \tag{4.25}$$

Which after the mean field approximation from [88]:

$$\frac{\mathrm{d}s}{\mathrm{d}m} = \frac{H}{N} \tag{4.26}$$

and

$$s = R - \frac{H}{2} \tag{4.27}$$

becomes:

$$F_s = k_b T \int_{stalk} \mathrm{d}m \frac{(\frac{H}{N})^{-1/2} a^{5/2} N_c}{(2R-H)L} = k_b T N \frac{(\frac{H}{N})^{-1/2} a^{5/2} N_c}{(2R-H)L} = k_b T N^{3/2} \frac{H^{-1/2} a^{5/2} N_c}{(2R-H)L} \tag{4.28}$$

Which after substitution for the number of chains in the brush N_c:

$$N_c = RL/d^2 \tag{4.29}$$

leads to a per chain excluded volume free energy of:

$$F_{ex} = k_b T (\frac{N^{3/2} a^{5/2}}{d^2}) \frac{RH^{-1/2}}{2R-H} \tag{4.30}$$

4.6.4 Free energy of cohesive interactions

Similar to the flat brush case we can define an energy density of blob interactions given that blob-blob interactions have an energy of $\epsilon k_b T$:

$$f_{coh} = 2\epsilon k_b T c^2 V \tag{4.31}$$

We have $c = N_c N_b / V$ the concentration of blobs (for N_b the total number of blobs per chain). For each chain or FG nup there exists only one sticky tip, therefore N_b =1 in this case. The volume the sticky tips can take on is approximated as an extended cylindrical region atop the stalk brush region whose volume is $V = 2\pi(R-H)L\delta$, with L equal to the height of the brush region axially along the pore.

N_c is the number of chains determined by the grafting distance d, with

$$N_c = 2\pi RL/d^2 \tag{4.32}$$

The concentration c is therefore:

$$c = \frac{R}{d^2 \delta (R-H)} \tag{4.33}$$

Which results in a free energy density of blob interactions of:

$$f_{coh} = 2\epsilon k_b T c^2 V \sim \epsilon k_b T \frac{R^2}{d^4 \delta^2 (R-H)^2} (R-H) L \delta \tag{4.34}$$

Which can be simplified to:

$$f_{coh} = \epsilon k_b T \frac{R}{d^2 \delta (R - H)} N_c \tag{4.35}$$

The total cohesive energy *per chain* is then equal to the volume of a sticky tip blob times the cohesive free energy density of all blob interactions divided by the number of chains.

$$F_{coh} = \epsilon k_b T \frac{R}{d^2 \delta (R - H)} N_c \delta^3 / N_c = \epsilon k_b T \frac{R \delta^2}{d^2 (R - H)} \tag{4.36}$$

4.6.5 Free energy for the total brush

We can now solve for the total free energy of the cylindrical brush per chain:

$$F_T = F_s + F_{ex} + F_{coh} \tag{4.37}$$

$$F_T = k_b T (\frac{(H/a)^{5/2}}{N^{3/2}}) + k_b T (\frac{a^{5/2} N^{3/2}}{d^2}) \frac{R H^{-1/2}}{2R - H} + \epsilon k_b T \frac{R \delta^2}{d^2 (R - H)} \tag{4.38}$$

$$= k_b T \left(\frac{(H/a)^{5/2}}{N^{3/2}} + \frac{a^{5/2} N^{3/2}}{d^2} \frac{R H^{-1/2}}{2R - H} + \epsilon \frac{R \delta^2}{d^2 (R - H)} \right) \tag{4.39}$$

4.6.6 Brush modeling parameters

We measured the β-CG Model R_g of Nsp1's sticky tip to be 1.8 nm and the R_g of its stalk to be 6.4 nm. For Nup100 we measured the β-CG Model R_g for the sticky tip to be 2.5 nm and its stalk to have an R_g of 2.5 nm. Free energy parameters N (blob number) and a (Kuhn length) were derived from the contour length L (number of AAs * 0.38 nm) and R_g values by solving the simultaneous polymer equations for an excluded volume chain [89]:

$$Na = L \tag{4.40}$$

$$0.377 N^{3/5} a = R_g \tag{4.41}$$

Estimation of the self interaction free energy for the sticky tips of the FG nups was derived from the ratio of the measured R_g in the β-CG Model of the sticky tips to the R_g of an ideal excluded volume chain of the same monomer length.

$$\Delta F_{self} = k_b T \left(\frac{R_g^{ex}}{R_g^{\beta}} \right)^2 \tag{4.42}$$

We estimated the self interaction energy of the Nsp1 sticky tip to be 4.7 kT. Similarly we estimated the self interaction energy of Nup100's sticky tip to be 8.0 kT. These self interaction energy levels set the energy scale for blob-blob interactions, which we refer to as ϵkT in the paper.

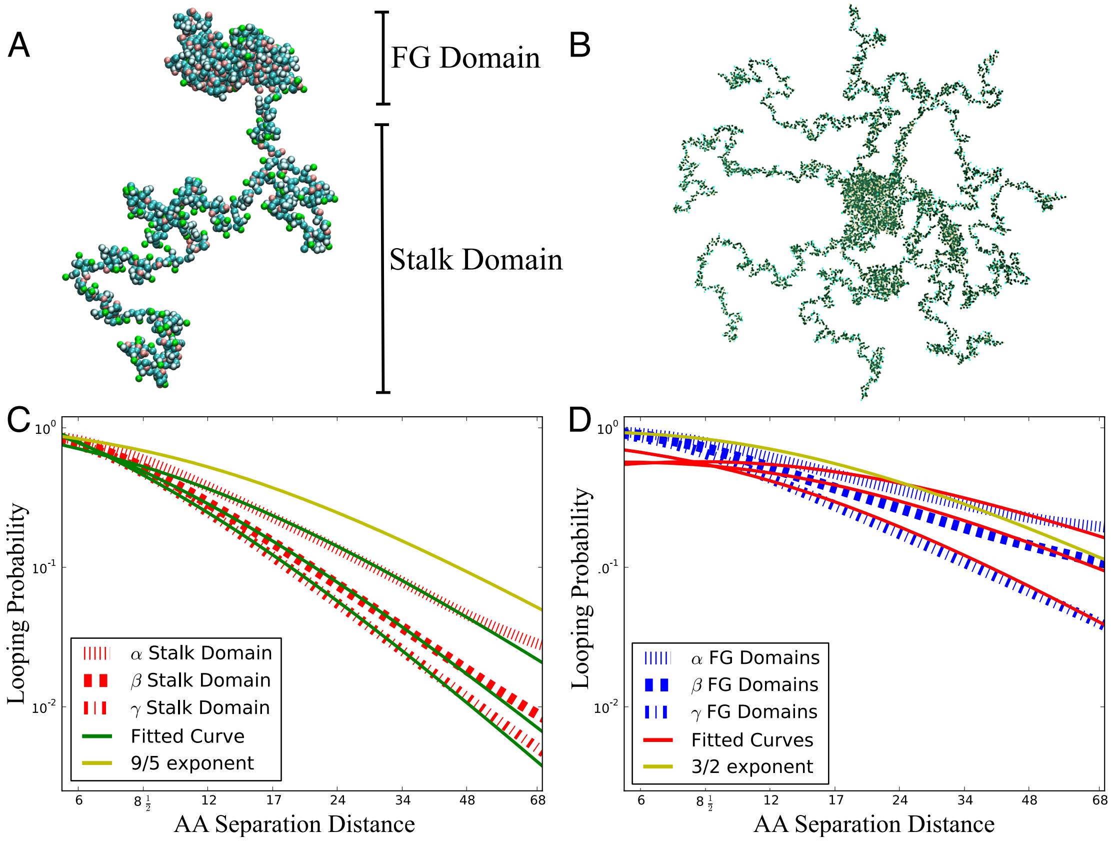

Figure 4.1: A) Simulation snapshot of full length Nsp1 showing the biphasic "FG Domain" and "Stalk Domain" structures in the α model. Snapshots for the other models β and γ are shown in Figure S7. (B) Snapshot of a ring of eight Nsp1's in the α model. (C) Log-log plot of amino acid contact probability (looping probability) as a function of the amino acid separation distance for stalk domains, averaged over simulated FG nups. The theoretical looping probability for extended coils (<1.6 nm cutoff) has an exponent of $p = 9/5$, and is shown in yellow. For the stalk domains the average fitted scaling exponents were 2.08, 2.58, and 2.82 for the different models α, β, and γ respectively. Stalk domains in these models had extended coil structures. (D) Similar average measured looping probabilities of the FG domains. For the α, β, and γ CG models the average fitted scaling exponents for the FG domains were 1.16, 1.38, 1.81 respectively. A theoretical looping probability with an exponent of 3/2 is shown in yellow, representing the dynamics of relaxed coil polymers. FG domains have looping exponents significantly lower than those of stalk domains, indicating the consistently large difference in ensemble structure between these two domain types. FG domains in the α model had a collapsed coil structure, a relaxed coil structure in the β model, and an extended coil structure in the γ model.

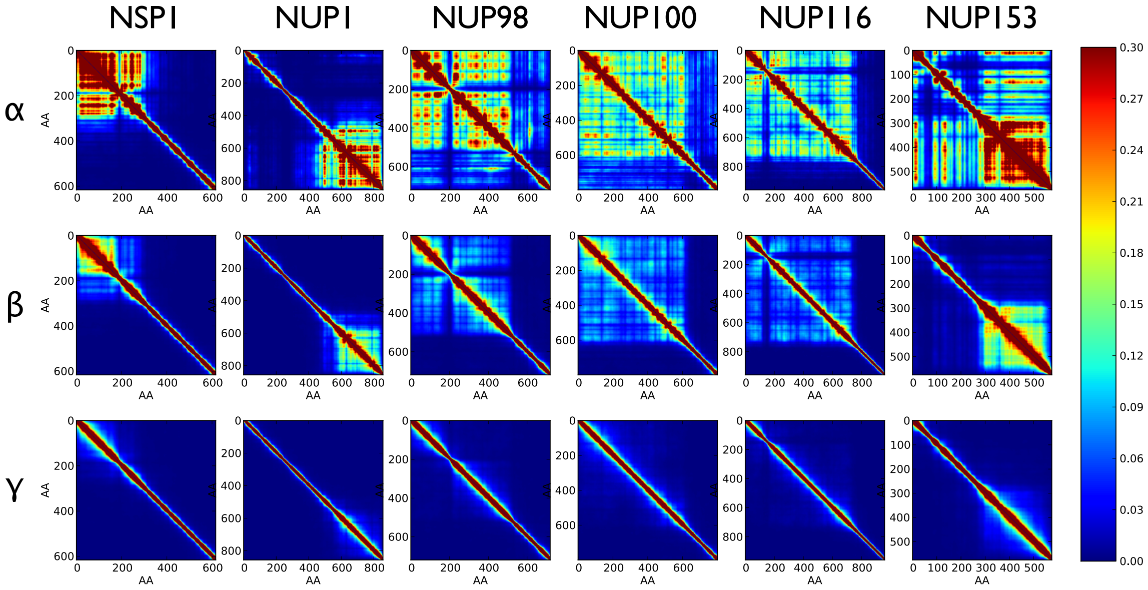

Figure 4.2: Contact maps for the different CG models α, β, and γ. Contact probability maps show the time averaged contacts between all pairs of amino acids. A block diagonal structure is noticeable in numerous contact maps, with one block for FG domains and diagonal contacts for stalk domains. A diblock structure can be most strongly seen in FG nups simulated under scenario β. Model α shows some block structure but often produces significant contacts between FG and stalk domains, as the two domains often interact. In striking contrast to models α and β, model γ produced nups with little difference in contact probability between FG and stalk domains with the entire FG nups representing unstructured extended homogenous polymers. Amino acid residues shown are with respect to the disordered domains of the simulated FG nups, while full protein AA indexes can be determined by domain definitions in Supplementary Material, Table 4.2.

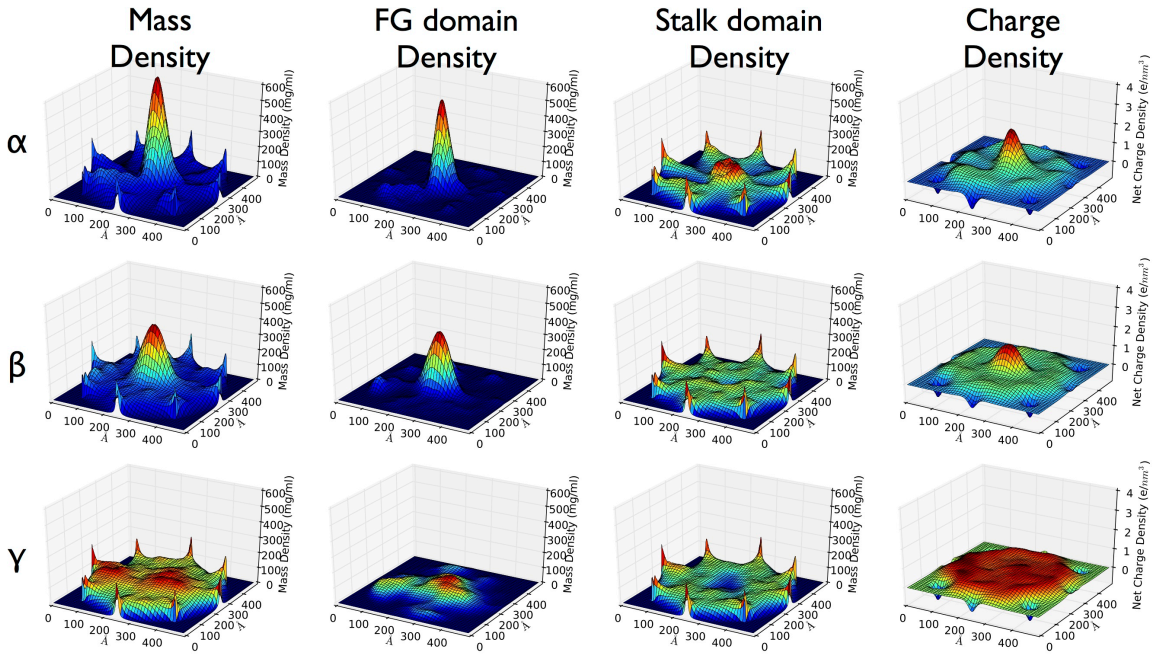

Figure 4.3: Density maps of a simulated pore containing 8 Nsp1 FG nups. Consecutively plotted for all three models is the total mass density (mg/ml), the mass density of "FG Domains" (mg/ml), the mass density of "Stalk Domains" (mg/ml), and the net charge density (e/nm^3). The total mass density clearly reveals a dense central plug connected by peripheral cables to the pore walls in the α and β models, while the γ model produces a homogenous extended brush. In the α and β models, one block phase (the FG domains) separates along the center of the channel while the other block (the stalk regions) aggregates along the periphery as can be seen in the spatial mass density for these domains. Positive charge density of the ring of Nsp1 nups mirrors closely the mass density of the FG domain, as FG domains are in general positively charged [81]. This produces a striking effect in the α and β models, with a strongly positively charged plug region forming which has a low density of charged amino acids (which are predominately located within stalk domain [81]).

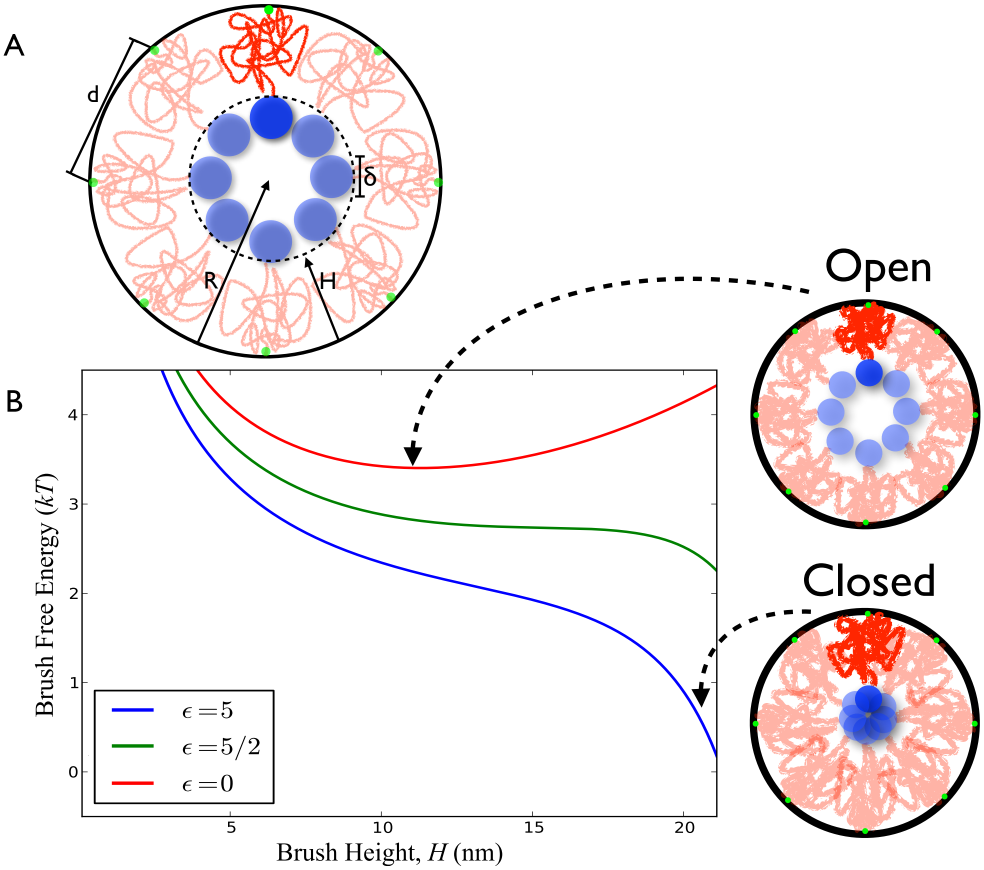

Figure 4.4: (A) Schematic of a polymer brush structure formed by diblock FG nups. Parameters H, height of the brush; R, radius of the pore; δ, diameter of the 'sticky tips'; and d, the average distance between anchor points. Green circles represent the locations at which FG nups are grafted to the pore. (B) Free energy of the Nsp1 brush as a function of brush height for various values of the blob cohesive energy ($\epsilon k_B T$). Brush heigh can extend to a maximum of around 22 nm, which is the radius R of the modeled pore minus the size of the sticky tips. Right: Schematic diagram of the proposed Diblock Copolymer Brush Gate model at various minima of the brush free energy. When particular transport factors are present which are able to outcompete the inter-FG domain "sticky tips" interactions, the brush is able to open up to a new free energy minimum that can accommodate the cargo. When interactions between sticky tips are able to recover into the several kT range, the pore is able to close with a free energy minimum at $H \sim R - \delta$. We estimated the self interaction energy level of the Nsp1 sticky tip to be 4.7 kT (Supplementary Material), which also sets the energy scale for blob-blob interactions of ϵkT.

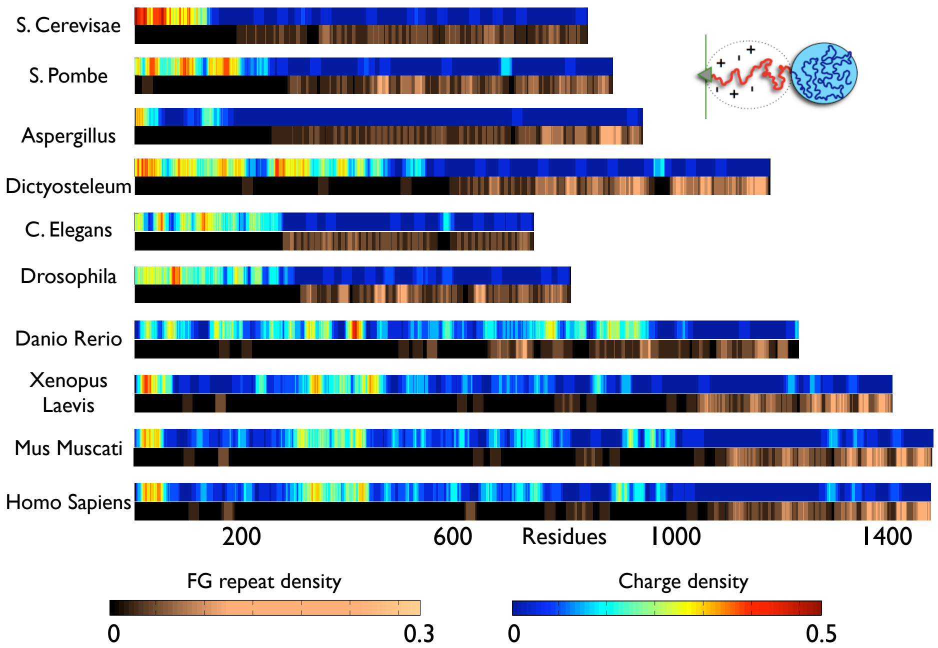

Figure 4.5: Heat maps showing FG repeat density (AA^{-1}) and charge density (AA^{-1}) across nucleoporins from ten different species. For each organism listed, the densities were measured along the disordered regions of the amino acid sequence of the FG nup which had the most FG motif repeats in that species. There appears to be a property common among all these heat maps, that regions high in FG motifs (pink) "FG domains" are in general disjoint from regions of the sequence high in charged amino acids (red, yellow, and light blue) "Stalk domains" [3]. Additionally each FG nup appears to conserve functional features of these domains, such as their orientation, and diblock polymer structure. For comparison, the FG density and motif locations for FG nups from S. cerevisiae can be seen in the Supplementary material, Figs. S8-9 The displayed FG nups are the ones with the most FG repeat per species, while each of these species also contain several other FG Nups whose structure does not fit this paradigm. Uniprot gene identifiers for each FG nup analyzed (from top to bottom) are Q02630, Q9UTK4, B0Y6T9, Q54EQ8, D1MN47, Q9VCH5, B8JIZ8, Q9PVZ2, Q80U93, and P35658 respectively.

4.6.7 Supplementary Figures and Tables

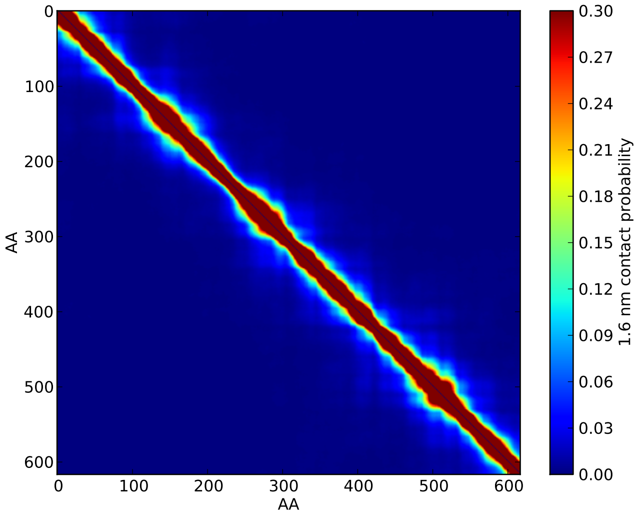

Figure 4.6: The contact probability map for a randomly shuffled Nsp1 sequence protein. Probability map of given AAs along the protein chain being closer than 1.6 nm during a 4 μs simulation for a randomly shuffled Nsp1 sequence in the β model shows no block diagonal structure or subsequent biphasicness.

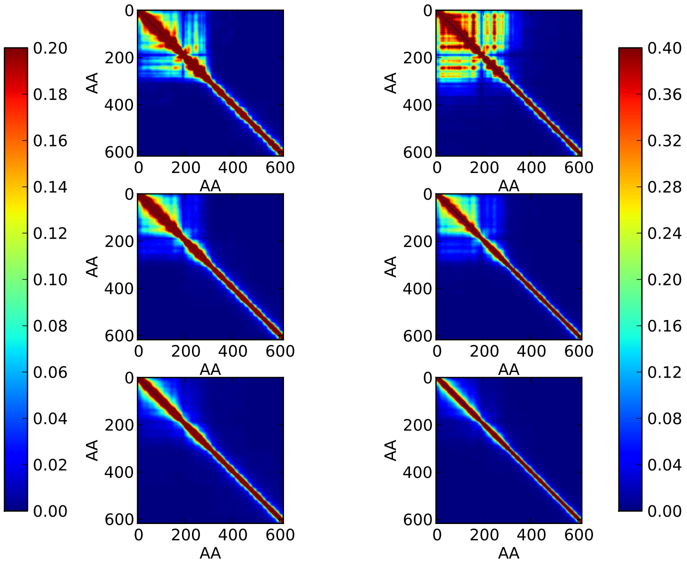

Figure 4.7: The contact probability map for individual Nsp1 sequences within ring structures (left) compared to free Nsp1 (right), with the top, center and bottom panels refering to the alpha, beta, and gamma models respectively. The average probability map of given AAs along the protein chain being closer than 1.6 nm during a 4 μs simulation for a Nsp1 sequence grafted to the inner walls of a cylinder as *in vivo* is similar to the contact map of Nsp1 free in solution. The contact map probability of individual free Nsp1s in solution has higher absolute contact probability due to a lack of interaction from other chains which can compete for contact. The contact map from the aggregate simulation demonstrates the same high level of dynamic movement of free Nsp1 even after confinement with partner FG nups in a cylindrical geometry.

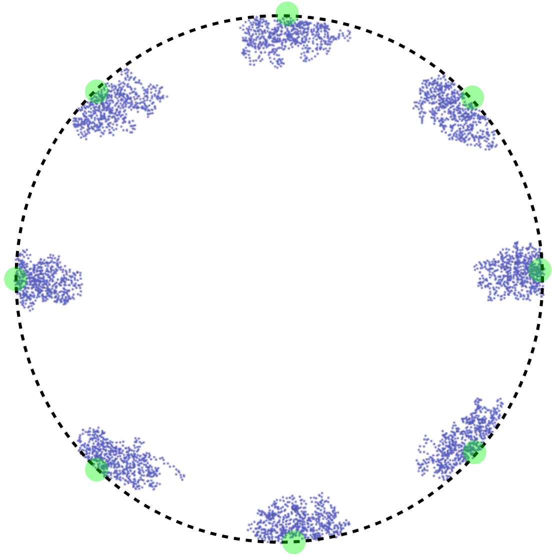

Figure 4.8: The ring of Nsp1 initial starting structure. Green circles represent the locations where 8 Nsp1 FG nups were grafted to the inner surface of a cylinder with the initial starting conditions for the ring simulations as shown. Initial positions of the Nsp1 FG nups was located near the wall of the cylinder, yet during simulations the ”sticky tips” of the FG nups aggregate towards the center of the cylinder in the α and β models.

Table 4.1: **Comparison of Model and Experimental R_g for human Nup153**

Model	Simulation	Experiment
α-CG Model	$R_g = 22.7 \pm 0.1\mathring{A}$	Collapsed Coil Observed [78], implied $R_g = 21.8 \pm 1.8\mathring{A}$ [79]
β-CG Model	$R_g = 45.8 \pm 1.3\mathring{A}$	–
γ-CG Model	$R_g = 57.9 \pm 3.4\mathring{A}$	Relaxed Coil Observed [80], implied $R_g = 54.0 \pm 4.6\mathring{A}$ [79]

Amino acid domain of human Nup153 simulated and analyzed for these comparisons was AAs 899-1475. Experimentally measured properties of disordered FG nups have yet to converge, therefore a number of CG models with different scaling factors (α, β, γ) were developed, with each scaling factor corresponding to a different class of experimental results. The CG model α was scaled such that it produces a collapsed coil (molten globule) [79] for the disordered region of nup153, as predicted by FRET measurements by Milles *et al* [78]. CG model γ was scaled such that it produces a relaxed coil [79] for the disordered region of nup153, as predicted by measurements on planar FG nup brushes [80]. Bead-halo experiments by Yamada *et al* (4) indicate nups are biphasic, with stalk domains non-interacting while FG domains are "sticky" to other FG domains. A scaling factor β precisely in the middle between α and γ was found to produce such biphasicness, with this scaled CG model labeled β.

Table 4.2: **Domain Definitions and Disordered Regions used**

FG Nup	Simulated Disordered Region	FG Domain	Stalk Domain
yNsp1	AAs 1-617	AAs 1-186	AAs 187-617
yNup1	AAs 220-1076	AAs 798-1076	AAs 220-797
hNup98	AAs 1-720	AAs 1-485	AAs 486-720
yNup100	AAs 1-800	AAs 1-610	AAs 611-800
yNup116	AAs 172-960	AAs 172-764	AAs 765-960
hNup153	AAs 899-1475	AAs 1195-1475	AAs 899-1194

Human FG nups are prefaced by an 'h', while FG nups from S. cerevisiae are prefaced with an 'y'.

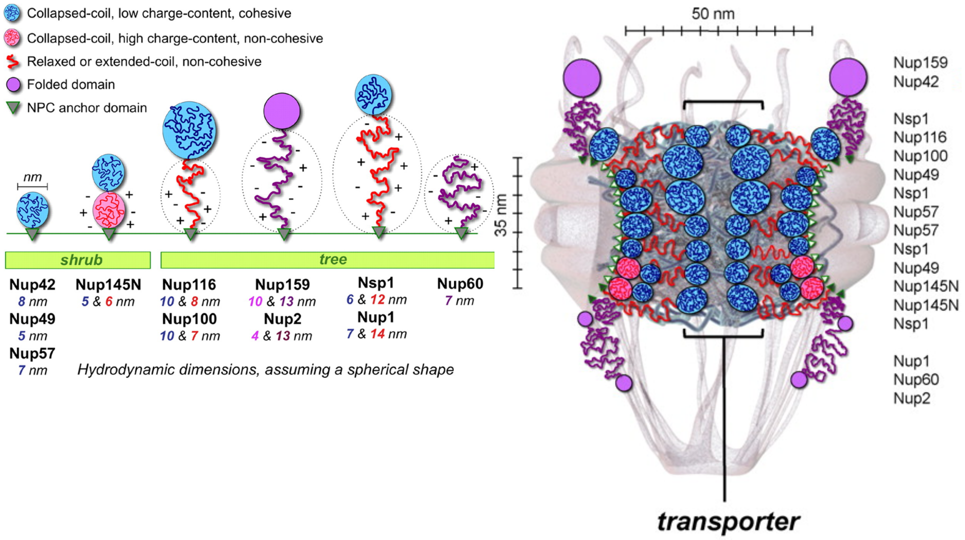

Figure 4.9: Left: Diagrams of natively unfolded regions of yeast FG nucleoporins and their hydrodynamic radii. Right: A diagram of the NPC architecture including the predicted topology and dimensions of yeast FG nucleoporins and their unstructured domains is shown. FG nucleoporins are listed according to the relative location of their C-termini along the z-axis, as determined by immuno-localization [3]. These figures and research were originally published in Molecular and Cellular Proteomics. Justin Yamada, Joshua L. Phillips, Samir Patel, Gabriel Goldfien, Alison Calestagne-Morelli, Hans Huang, Ryan Reza, Justin Acheson, Viswanathan V. Krishnan, Shawn Newsam, Ajay Gopinathan, Edmond Y. Lau, Michael E. Colvin, Vladimir N. Uversky, and Michael F. Rexach. A Bimodal Distribution of Two Distinct Categories of Intrinsically Disordered Structures with Separate Functions in FG Nucleoporins. *Molecular and Cellular Proteomics.* 2010; 9:2205-2224. © the American Society for Biochemistry and Molecular Biology.

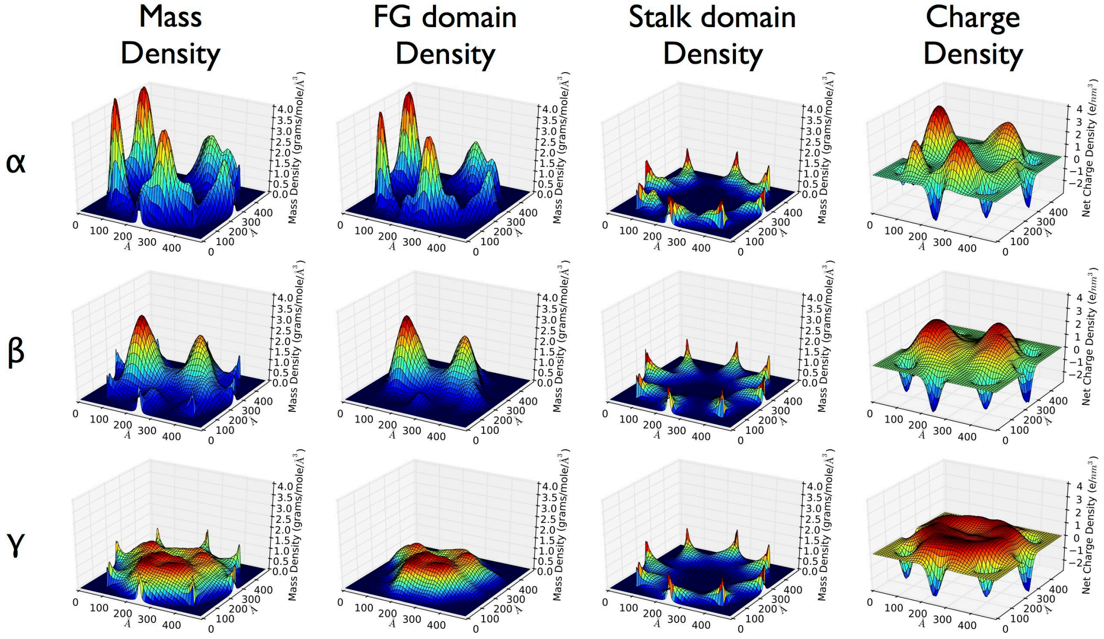

Figure 4.10: Density maps for a ring of 8 Nup100s in different modeled scenarios. Consecutively plotted for all three models are the total mass density within a simulated pore (mg/ml), the mass density of "FG Domains" (mg/ml), the mass density of "Stalk Domains" (mg/ml), and the net charge density (e/nm^3). The total mass density reveals that a dense central plug is unable to form in the α and β models, leaving the pore essentially open. On the other hand the γ model produces a homogenous extended brush resulting in a filled in and closed pore. Positive charge density of the ring of Nup100 nups mirrors closely the mass density of the FG domain, as FG domains are in general positively charged (4).

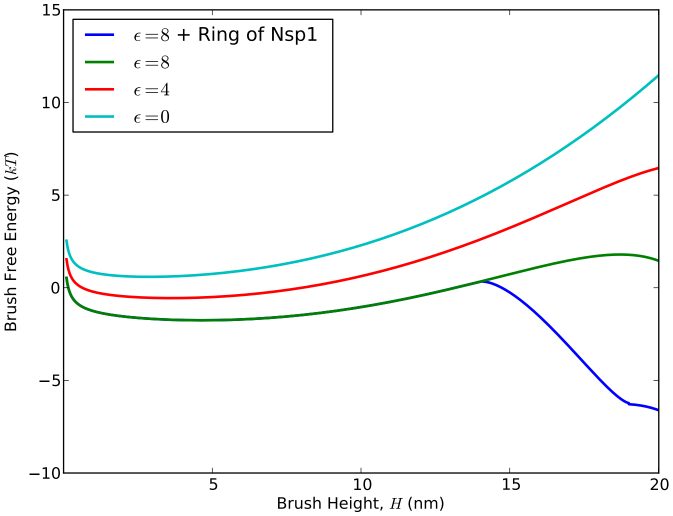

Figure 4.11: Free energy of the Nup100 brush as a function of brush height for various values of the blob cohesive energy ($\epsilon k_B T$). Brush heigh can extend to a maximum of around 20 nm, which is the radius R of the modeled pore minus the size of the sticky tips. When particular transport factors are present which are able to outcompete the inter-FG domain ”sticky tips” interactions, the brush is in an open state (light blue). When interactions between sticky tips is able to recover into the several kT range and a closed ring of Nsp1 is adjacent to the Nup100 ring, the pore is able to close with a free energy minimum at $H \sim R - \delta$. We estimated the self interaction energy level of the Nup100 sticky tip to be 8.0 kT, which also sets the energy scale for blob-blob interactions of ϵkT.

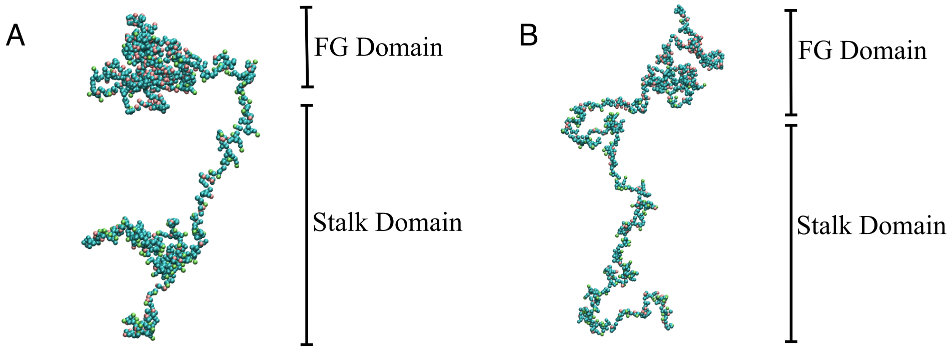

Figure 4.12: Simulation snapshots. A) A simulation snapshot of full length Nsp1 showing the biphasic "FG Domain" and "Stalk Domain" structures in the β model. B) A simulation snapshot of full length Nsp1 snapshot as in part A) for the γ model, while the α model snapshot is shown in the main text.

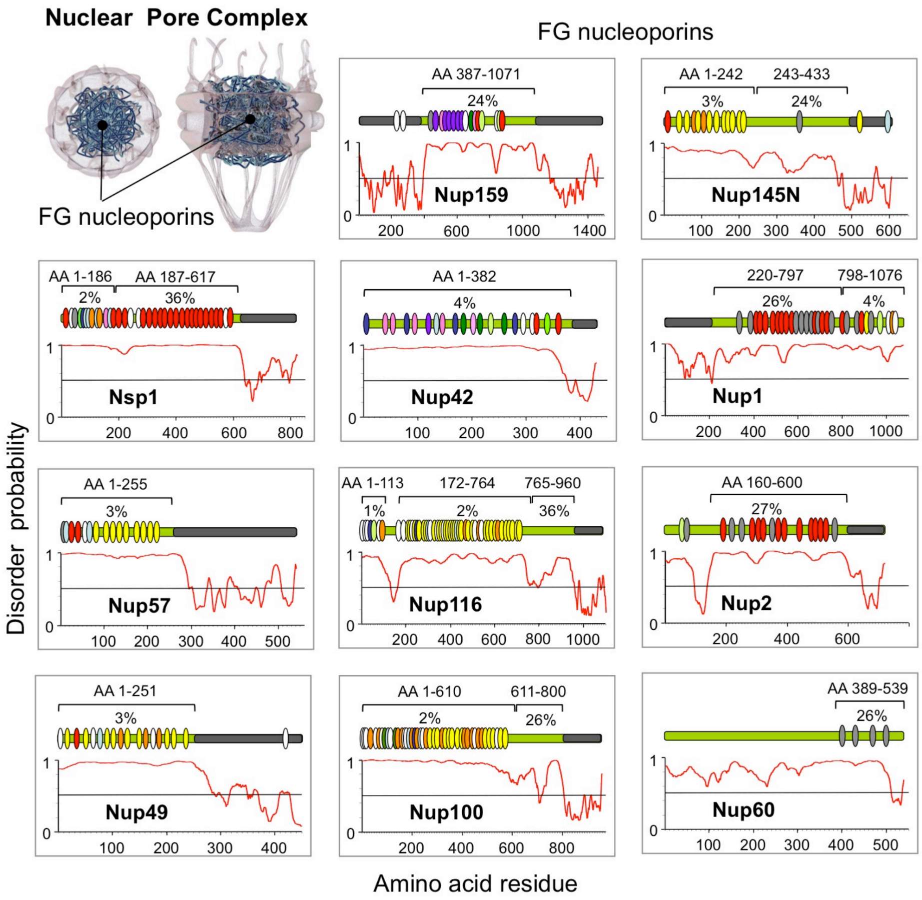

Figure 4.13: FG repeat locations among FG nups. The locations of FG motifs and their type are shown for FG nups in S. cerevisiae. GLFG motifs are colored yellow, FxFG red, SPFG dark green, FxFx light gray, SAFG dark blue, PSFG bright green, NxFG light blue, SLFG orange, xxFG white, FxxFG lime green, double FG motifs (SAFGxPSFG) are pink, and the triple FG motifs are purple. This figure was originally published in Molecular and Cellular Proteomics. Justin Yamada, Joshua L. Phillips, Samir Patel, Gabriel Goldfien, Alison Calestagne-Morelli, Hans Huang, Ryan Reza, Justin Acheson, Viswanathan V. Krishnan, Shawn Newsam, Ajay Gopinathan, Edmond Y. Lau, Michael E. Colvin, Vladimir N. Uversky, and Michael F. Rexach. A Bimodal Distribution of Two Distinct Categories of Intrinsically Disordered Structures with Separate Functions in FG Nucleoporins. *Molecular and Cellular Proteomics*. 2010; 9:2205-2224. © the American Society for Biochemistry and Molecular Biology.

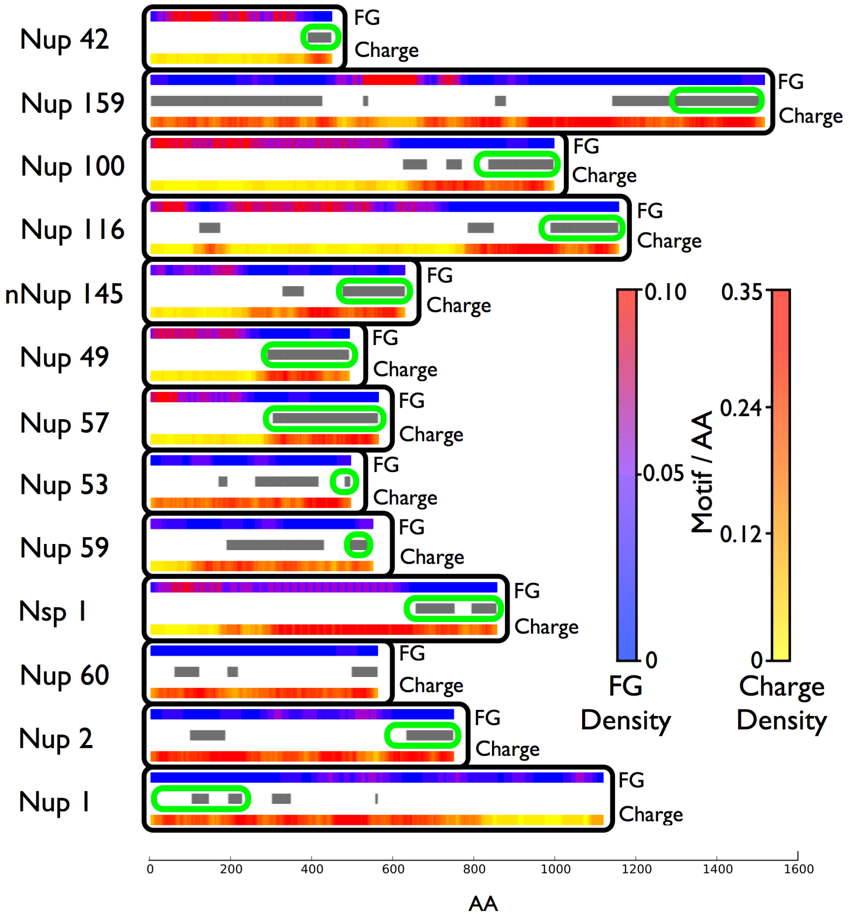

Figure 4.14: FG repeat density among FG nups. The spatial distribution of FG motifs and charged AAs for all known FG nups of S. cerevisiae plotted as motif/AA, averaged over 20 nearest AAs. Regions of high FG motif density are shown in pink while regions of low charge density, shown in yellow, roughly correspond spatially throughout the sequences of these nups. Regions of protein which are predicted to form folded structures by the PONDR algorithm are highlighted with grey bars, and known/predicted [3, 44] anchor domains circled with green ovals. This figure was originally published in *PLoS one*. Ando, D., M. Colvin, M. Rexach, and A. Gopinathan, 2013. Physical Motif Clustering within Intrinsically Disordered Nucleoporin Sequences Reveals Universal Functional Features, licensed under CC BY 2.5.

Chapter 5

THE DIBLOCK COPOLYMER BRUSH GATE MODEL WITH ”SHRUBS”

5.1 Introduction to ”Shrubs”

My work so far in modeling of FG nups and the NPC has ignored the so called ”shrub” FG nups. In a previous analysis by Yamada *et al* [3], it has been reported that FG nups from Baker’s yeast are present in a bi-modal distribution *in vivo*, with some FG nups being di-block ”trees” while others are single-block ”shrubs”, Fig. 5.2A. The di-block tree FG nups have one disordered region near the grafting end which is extended and contains a high density of charged amino acids and another disordered region at the free end which is collapsed and contains numerous FG repeats and relatively few charged amino acids. These bi-phasic ”tree” FG nups form a copolymer brush which fills the NPC in the DCBG model of transport (Fig. 5.1A), where the approach of certain forms of cargo can alter the local brush structure, effectively opening the closed brush structure and allowing for transport. In the Forest model of NPC structure by Yamada *et al* [3] Fig. 5.2B, single-block FG nups localize to the pore wall, in addition to the di-block FG nups modeled by the DCBG model. Deletion experiments on single-block FG nups have also revealed that they are critical for proper functioning of the NPC in Baker’s yeast [60,92]. Here we attempt to understand what role single-block FG nups may play when introduced to the DCBG model, Fig. 5.1B, which we refer to as our modified DCBG model.

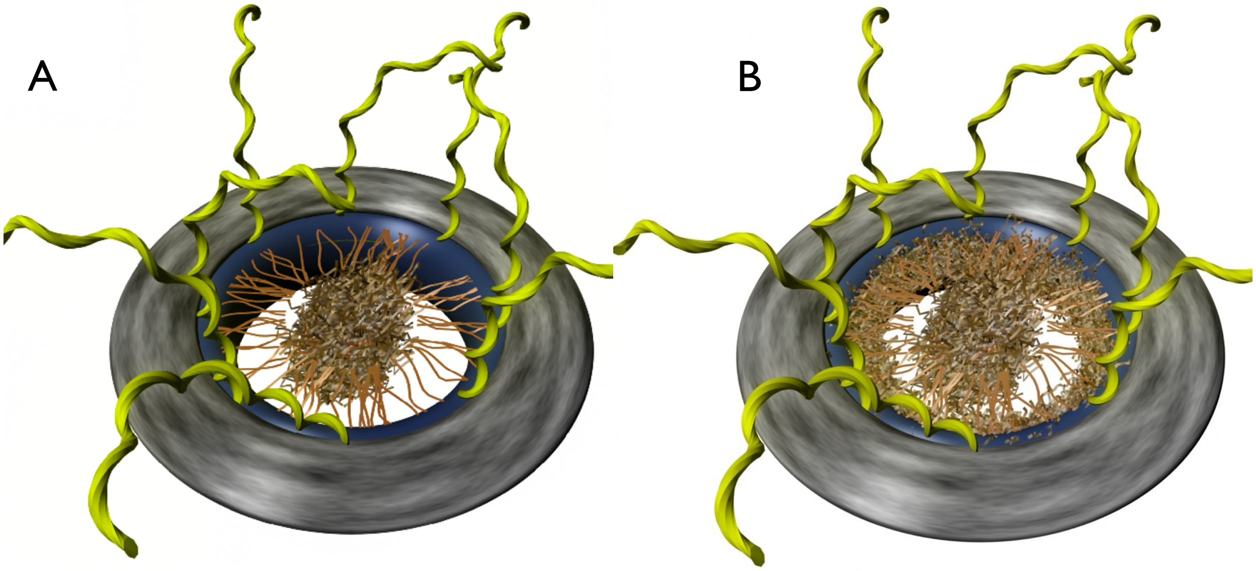

Figure 5.1: Illustration of DCBG models. Each NPC depicted spans the double lipid bilayer nuclear envelope, oriented such that the tops of the pores face the cytoplasm, while the bottoms of the pores face the nucleus. (A) In the DCBG model individual di-block FG nups have collapsed coil gel-like regions and extended coil brush-like domains, resulting in a microphase separation of these domains within the NPC. This results in a central plug-like structure supported by a polymer brush of extended disordered regions of FG nups. [112] (B) The modified DCBG model is the same as part A except for the addition of a dense region of single-block FG nups which lies along the wall of the NPC.

Although the NPC plays a key role in eukaryotic biology and numerous studies have been performed on the structure and properties of individual nucleoporins [4,5,8,15–18,28], the structure of the polymer complex within the NPC and its actual mechanism of transport regulation remain unclear. Different models for NPC structure and function, such as the "hydrogel" [12] and "virtual gate" [14] models, assume very different morphologies for the polymer complex of disordered proteins which fills the nuclear pore. For example, the hydrogel model predicts that FG nups interact via hydrophobic amino acids to form a dense filamentous meshwork that physically blocks protein diffusion while the virtual gate model predicts that FG nups have limited hydrophobic interactions and instead take on an extended polymer brush like structure that form an entropic gate at the NPC which blocks protein diffusion. Another model, the *Di-block Copolymer Brush Gate* (DCBG) model [112], Fig. 5.1A, assumes that the key FG nups which regulate transport are individually bi-phasic, while most theoretical and polymer physics approaches [19–22] to the

NPC transport problem tend to assume a homogenous structure for individual FG nups, resulting in a relatively homogenous NPC architecture.

Results from our molecular simulations demonstrate that single-block FG nups are compact and therefore localized to their anchor points within the NPC. These anchor points are located on the inner wall of the NPC [113]. Through polymer brush modeling we find that adding a layer of single-block FG nups to the DCBG model increases the range of cargo sizes which are able to translocate the pore via a cooperative interaction with the tips of di-block FG nups. This suggests a functional reason for the presence of both single-block and di-block FG nups *in vivo*, with the combination increasing the diameter of cargo which can translocate the pore. This cooperative mechanism also provides a framework for enhancing polymer brush based pore transport in biomimetic applications.

5.2 Results

5.2.1 Coarse Grained Simulations

We first examine the polymer conformations adopted by single-block FG nups using molecular dynamics (MD) simulations. Given the large size of disordered of FG nups and the long timescales involved in their dynamics, coarse grained molecular dynamics simulations were utilized. The coarse grained (CG) model by Hills *et al.* [27] was used, which is detailed enough that it is able to accurately reproduce the secondary and tertiary structure of proteins, and maintains the ability to fold small proteins correctly *ab initio* starting from fully extended states. An implicit solvent is used resulting in rapid simulation speed and which allows for the capability of simulating multi-protein systems over microsecond timescales [112]. The Hills model reproduces the energy landscape of proteins as found *in vivo* by representing each amino acid by multiple CG bead interaction sites and by using specially developed interactions between these sites. These interactions between these CG beads are determined by measuring the forces present in highly accurate all atom simulations of diverse protein sequences to create a CG interaction following the multiscale coarse-graining (MS-CG) method [75]. This multiscale method makes no a priori assumptions about the form of interactions between CG beads, and implicitly includes multibody correlations [76] in the resulting effective CG potentials. The secondary and tertiary structure of proteins emerges naturally in the CG simulations through the comprehensiveness of the physical model, with the emergent CG model interactions general enough to simulate proteins of arbitrary amino acid sequence.

CG models typically require a normalization step where CG potentials for a specific protein simulation are scaled such that model results match well with experiment [27, 77]. In a previous publication [112] we found a scaling factor for the Hills model which reproduces experimental results obtained by Yamada *et al* [3]. These experiments indicate that many FG nups are individually di-blocks, with one domain resembling an extened polymer and non-interacting in general while

FG domains are collapsed coils and cohesive to other FG domains. Our previously determined scaling factor reproduces the biphasicness of FG nups in terms of both cohesion and polymer structure. The FG nups simulated in this manuscript are of the single-block type from *S. cerevisiae*, consisting of Nup42, Nup49, and Nup57. The amino acid regions of the FG nups simulated and domain definitions used are from Yamada *et al* [3], therefore our simulations of single-block FG nups can be directly compared to their experimental characterizations.

The single-block FG nups simulated in this paper have a single block amino acid sequence structure which can be described as a single FG domain, corresponding to a continuous sequence region with a low concentration of charged amino acids and a high concentration of FG motifs [81]. Polypeptide chains representing the disordered regions of these FG nups were simulated from an initially fully extended configuration and equilibrated before the start of simulation for 1 μs. Fig. 5.3A shows a typical snapshot of one such nup, Nup57, after collapse and equilibration. This snapshot indicates that single-block FG nups tend to exhibit a single block polymeric structure which resembles a collapsed coil structure. Nup42, Nup49, and Nup57 had R_g values of 2.4 nm, 2.0 nm, and 2.0 nm respectively over 4 μs of simulation which is consistent with these FG nups being collapsed coils.

To characterize the polymeric structure of individual FG nups, we defined the contact probability between a pair of amino acids as the percentage of time the corresponding CG backbone beads are within 16 Angstroms of each other. This contact probability is plotted as probability maps over all possible amino-acid contact pairs, Fig. 5.3B and Fig. 5.4. Generally, these contact maps show a monolithic block diagonal structure for the single-block FG nups that represents a collapsed disordered structure, except for a very short extended domain where Nup42 is anchored to the NPC pore wall. For all simulated nups, the contact maps indicated that the individual FG nups were highly dynamic and disordered. This can be seen in the relatively smooth block contact structure of simulated nups, with contacts between amino acids rapidly exchanging and no folded constant contacts visible. The contact maps of our simulated single-block FG nups differ significantly from previously reported contact maps of di-block FG nups (Fig. 5.3B) [112]. Single-block FG contact maps generally lack an extended disordered domain which predominately avoids contact with itself and FG domains. Here, our simulation results on the full disordered regions of single-block FG nups support the hypothesis that FG nups exists in a bimodal configuration [3] by demonstrating that it is clearly possible that individual FG nups can form ”shrub” like single-block polymers rather than ”tree” like di-block structures which have been reported earlier [3].

5.2.2 Polymer brush morphology and dynamics with single-block FG nups

We turn to polymer physics modeling to understand the underlying biophysical mechanisms that may govern the formation and dynamics of different types of mesoscale structures within the NPC which play a role in transport regulation. First, we summarize a simple polymer brush model, the *Di-block Copolymer Brush Gate* (DCBG) model [112], that has been successfully used to model aggregates of di-block type FG nups. This model treats the core of the NPC as an idealized cylindrical polymer brush constructed of di-block polymers. It is assumed that the extended domain region of the nups can be described by neutral excluded volume polymers with an appropriate length and stiffness, and that they are grafted to the inside of a cylinder that is geometrically representative of the nuclear pore. The biphasic structure of these gating di-block FG nups is incorporated by modeling the ends of the polymers, the FG domains, as structureless cohesive blobs that correspond to the cohesive FG domains at the FG nup tips. The resulting configuration is a polymer brush as shown schematically in Fig. 5.1A and Fig. 5.5. Here we extend this model to include the presence of single-block FG nups around the pore wall, which are present in Baker's yeast, Fig. 5.1B [3].

We consider a polymer brush confined to the inside of a cylinder of radius R, with di-block FG nups modeled as chains of length N Kuhn lengths (a) with sticky tips of size δ with a grafting distance d, see Fig. 5.5. An equal number of single-block FG nups were modeled as sticky blobs of size δ_s grafted along the pore wall. In a modified DCBG model with single-block FG nups present we can solve for the height H of the resulting brush via mean field polymer brush modeling. The height of this polymer brush at equilibrium can be derived via a minimization of the free energy of the brush which has contributions coming from (i) entropic stretching of the extended domains (ii) the excluded volume interactions (steric hindrance) between the extended domains (iii) the cohesive energy interactions between the tip blobs and (iv) the cohesive energy interactions between tip blobs and single-block FG nups. The modified DCBG model includes an extra term (iv) which accounts for the attraction of the single-block FG nups with the tips of di-block FG nups. The DCBG model and its single-block FG nup extension, which are both models of polymer brushes over a negatively curved surface with cohesive tips coupled to the free ends of the polymers, result in a brush height which is given by the solution to a non-linear differential equation. An analytical solution however can be arrived at by making a mean-field approximation, which results in an algebraic form for the effective free energy per chain [112];

$$\frac{F}{k_BT} \sim (H/a)^{5/2}N^{-3/2} + \frac{a^{5/2}N^{3/2}RH^{-1/2}}{d^2(2R-H)} - \frac{\epsilon\delta^2 R}{d^2(R-H)} - \frac{\epsilon_s\delta^2(\delta_s-H)R\Theta(\delta_s-H)}{d^2\delta_s R_s}$$

The Heaviside step function $\Theta(x)$ is defined to be 1 if x is positive and 0 if

x is negative, with $\epsilon k_B T$ and $\epsilon_s k_B T$ equal to the effective cohesive energy between a pair of tip blobs and between single-block FG nups and di-block FG nups tip blobs respectively. The first two terms of the free energy follow directly from prior work on polymer brushes in cylinders [88, 89]. The Kuhn length a and the effective polymerization number N in the brush free energy have been measured to be twice the monomer length of amino acids in the chain and half the number of amino acids in the disordered protein chains, respectively [112]. R_s is the radial distance at which the single-block FG nups are located at within the pore. The cohesive domains at the tip are in a collapsed state as measured from CG modeling [112], with the sticky tip blobs of size (diameter) $\delta = 3.6$ nm, twice the radius of gyration as measured previously via CG model simulations [112]. We make the assumption that the average spacing between grafting points of a given di-block FG nup, for example Nsp1, along the pore wall is $d \sim 10$ nm, which we estimate by assuming that there are 32 copies [5] of Nsp1 anchored along the inner wall in a symmetric fashion along the geometry used in Yamada et al [3]. The pore radius is taken to be $R \sim 25$ nm [3]. For the single-block FG nups size δ_s we use a value of 4 nm which is roughly twice the average radius of gyration of collapsed FG domains of single-block FG nups Nup42, Nup49, and Nup57 which are relatively similar in size. The location of the single-block FG nups are not precisely known, with $R_s = 25$ nm in scenarios where the single-block FG nups are located along the pore wall [113]. Term (iv), the cohesive energy interaction between tip blobs and single-block FG nups was derived from the probability of single-block FG nup and tip contact, which equals the product of tip and single-block FG nup concentrations within the pore as they spatially overlap, while the cohesive energy interactions between the tip blobs was similarly derived as a tip to tip contact probability, equal to the square of the tip blob concentration as derived [112].

Fig. 5.5 shows the free energy of the modified DCBG model as a function of the brush height for three different values of the tip-tip cohesive energy and a fixed tip to single-block FG nup cohesiveness of $\epsilon_s = 6kT$. For a given tip-tip cohesive energy, the equilibrium value of the brush height is located at the minimum of the free energy profile. When there is no tip cohesion, corresponding to situation where there is complete screening of tip-tip interactions by transport factors and $\epsilon = 0$, we can see that the equilibrium brush height is lowered to less than at least half the pore radius, which implies that if cohesion between the tips is lost a wide-open channel down the center of the pore can be formed for cargo to travel through. We term this the "open" state of the gate. As the cohesive energy is increased, we see a new minimum developing at $H \sim R - \delta$, which is a state where the extended domains have stretched to such a degree that the cohesive tips are at the center of the pore, completely closing the conduit. We term this as the "closed" state of the pore. This indicates that there can be a transition between the closed and open states of the polymer brush as the tip-tip cohesion energy is varied. Thus the arrangement of individually biphasic functional polymers, with an extended domain and collapsed cohesive tip, leads to a brush structure that can switch between open and closed states, as shown in Fig. 5.5, with relatively small changes in the cohesiveness of polymer tips. Importantly, the free energy of the brush indicates that the nups extend all the way into the center of the pore when transport factors are not present and tip-tip cohesion is high, while the FG nups form a brush with an intermediate height that leaves an open transport conduit for transport along the middle of the pore when tip-tip cohesion is low.

With regards to the function of gated NPC transport *in vivo*, it is known that transport factors, such as karyopherins and exportins, bind to FG motifs [90, 91]. This provides a mechanism which can screen tip-tip interactions, lowering ϵ via the disruption of the inter-molecular cohesiveness between nup tips by the competitive binding of transport factors. We propose that the binding of transport factors *in vivo* results in such a switch from the closed to open state forming the basis for our DCBG model. Specific types of transport mediated cargo complexes, which have an attraction to di-block FG nup tips, could therefore have the ability to dissolve into the central plug that keeps the gate closed, thus effectively opening it up to the desired diameter. Once open, these cargo complexes can then undergo one dimensional diffusion down the center of the NPC channel, until enzymes on the nuclear side cleave the transport factors which reduces the propensity of cargo to remain in the channel leading to a rapid exit to the nucleus. For export, this picture would be reversed with cargos starting in the nucleus and cleaving enzymes located on the cytoplasmic side of the NPC. From the free energy of the DCBG and its modified model, we clearly see that the presence of single-block FG nups increases the equilibrium opening of the NPC, Fig. 5.6A, regardless of the single-block FG nup anchor location within the pore. Thus, the presence of single-block FG nups therefore allows the pore to open to a considerably greater extent.

We also used polymer brush modeling to understand in detail how the strength of the effective tip-tip cohesion, which is modulated by the presence of transport factors, changes brush structure and its receptivity to transport, Fig. 5.6A. We find that tip-tip cohesiveness above ∼1.9 kT (when single-block FG nups are present) provides for a critical level of tip cohesion which allows the pore to be in a closed state. Lower tip cohesiveness results in a rapid transition in brush structure, with the brush opening up to 40 nm in diameter below the critical tip cohesiveness. This rapid change in brush structure as a function of tip cohesion is a result of the relatively flat free energy of the brush system, Fig. 5.5, where large changes in brush height can have small differences in the brush free energy. This is likely to be functionally useful in designing generalized "spring-loaded" polymer gating pores.

Similar to how transport factors and cargo screen tip-tip interactions to modulate tip-tip cohesiveness, transport factors also introduce a free energy of binding to tips into the pore brush system which allows for work to be done onto the brush. Given the free energy profile of the brush, work done on the brush typically results in a lowering of brush height and an opening of the pore, Fig. 5.5. The maximum amount of brush opening assuming all the introduced tip-cargo free energy is converted into work onto the brush is plotted in Fig. 5.6B. This brush opening is plotted relative to different levels of screening of tip-tip interactions. Different degrees of screening will result in different effective tip cohesiveness which are plotted over possible values of 3 kT and 6 kT as cargo move through the pore. For any given fixed tip-tip interaction, there exists a sharp transition from closed to open brush states as the transport factor-tip binding energy as increased, beyond which the return from increasing the binding energy further results in negligible increased pore opening. Again, this is a direct result of the relatively flat free energy curve of the brush. For values of $\epsilon \sim 3\ kT$ the maximum brush opening for a given amount of work done on the brush can be up to four times greater, Fig. 5.6B, if single-block FG nups are present. The localization of single-block FG nups along the pore wall therefore results in a transition from closed to open brush states at lower free energies in both tip-tip cohesion and tip-cargo interactions.

5.3 Discussion

The mechanisms behind NPC transport regulation remain poorly understood, and many experimental results regarding the basic properties of individual FG nups such as their structure, location and cohesive properties are in conflict. The aggregate properties of these FG nups when put together within the confines of the NPC channel is even more uncertain. Here we have attempted to understand how the overall architecture of the assembly of both ”tree” type di-block FG nups and ”shrub” type single-block FG nups in the NPC channel is governed by the properties of the individual FG-nups and their physical polymer properties. Using a model where individual FG nups can have specific domains which are compact or extended, our results indicate that a Forest type model structure [3] emerges, with the ”sticky tips” of di-block FG nups coalescing in the center of the pore to form a hydrophobic plug, which is connected by an extended coil brush zone to the pore wall where single-block FG nups are localized. This three part structure results in a spring loaded polymer brush gate which is capable of opening for very large cargo.

In this paper we have focused on the effects of adding single-block FG nups to the larger di-block FG nups, both of which have been shown to be critical for many forms of transport [92] in Baker’s yeast. Some details of NPC structure such as cytoplasmic filaments and nuclear basket structure have not been considered in our analysis. Since the cytoplasmic filaments and the basket are spatially well separated from the interior of the pore we anticipate that they would not interfere with the copolymer brush structure. Given that the properties of the single-block FG nups are similar to the sticky tip FG domains of di-block FG nups in terms of bimodal adhesion [3], we show that the single-block FG nups interactions with di-block FG nup tips helps to stabilize the wide open configurations of the copolymer brush in the presence of large cargo in the middle of the pore. This sort of cooperativity could be relevant for the export of large ribonucleoprotein particles whose transport is facilitated by surface bound transport factors. In the work of Strawn *et al* [92] single-block FG nups like Nup49 or Nup57, together with di-block FG nups like Nup100 or Nup116, were required for the viability of yeast cells, yet single single-block FG nup or single di-block FG nups deletions resulted in viable cells. Given that the formation of the NPC's permeability barrier is the most fundamental function of the NPC, this deletion data likely indicates that the presence of both classes of FG nups are required for the NPC in Baker's yeast to spatially form a permeability barrier across the pore using a central plug along the core and "shrub" FG nups along the wall. Additionally, deletions of just Nup49 or Nup57 resulted in only negative perturbations to large cargo mRNA export, while deletions of di-block FG nups only had no effect on mRNA export [60], indicating that single-block FG nups are likely necessary for the transport of large cargoes via cooperativity with the sticky tips of di-block FG nups. Our modified DCBG model can explain both the requirement for having both single-block FG nups and di-block FG nups for viability and the necessity of having a critical number of single-block FG nups present for the transport of large cargos such a mRNA.

We have studied the equilibrium structure of the NPC brush for various tip-transport factor interaction energies Fig. 5.6B. Tip-tip interaction energies and the level of screening by transport factors remains unknown, therefore resultant optimal tip-transport factor interaction energies for brush opening remain unknown. Qualitatively we find that as the level of tip-tip screening by transport factors increases, the free energy gained from tip-transport factor binding that is required to open the pore decreases strongly. Given that rapid cargo transport through the NPC is likely due to rapid exchange of transport factor-tip partners as the cargo moves through the pore, a high off rate equivalent to the inverse of the tip-tip exchange timescale will be advantageous. A high off rate such as this necessarily implies that the tip-transport factor binding energy will be small. With a small tip-transport factor binding energy, opening of the pore must rely more on transport factor screening of the di-block FG nup tips rather than thorough the free energy change of the brush through tip-transport factor binding. The optimal properties of transport factors are therefore likely to involve two factors, low tip binding energies for a high off rate and a high propensity for tip screening via geometric or competitive binding effects. Additionally, maximizing the total cargo flux likely involves simultaneously opening the pore around as many cargoes as possible while maximizing the transport factor off rate. We find that both increased tip-tip screening or increased tip-cargo binding energies can achieve increased brush opening (Fig. 5.6), although increased tip-tip screening by transport factors is more desirable as the off rate is less likely to be reduced, resulting in more rapid transport.

Our results not only provide for a consistent physical mechanism by which we can understand nucleocytoplasmic transport and the puzzling presence of ”shrub” type single-block FG nups, but our model is also very much applicable to designing and optimizing novel forms of biomimetic transport. The application of biomimetic membranes [102] has a wide variety of uses ranging from chemical and biological separation to purification [103]. Additionally these membranes provide a platform for the analytical detection of substances, drug delivery [104], as well as for self-contained reactors and mock cells [105, 106]. The regulated cross-membrane trafficking of cargos through the membranes in such systems is of high importance and our investigations of a relatively simple biologically compatible mechanism could be very useful. In general, any di-block polymers with the extended domain near the polymer grafting point and a collapsed and cohesive tip domain could be used to design biomimetic pores which regulate traffic using our DCBG mechanism. Similar to our modified DCBG model, cohesive collapsed coil polymers can be added along the biomimetic pore wall to increase the size of cargos which can transport the pore. As evidenced by early efforts to make biomimetic gates inspired by the NPC [107, 108], our DCBG models should find many opportunities for implementation outside of the NPC context from which they are inspired.

5.4 Methods

CG simulations were run using the model of Hills *et al* [27] in the large-scale atomic/molecular massively parallel simulator (LAMMPS) [109] software package. Using a scaling factor of 4.30 at 300 K, simulated single-block FG nups from *S. cerevisiae* were equilibrated for 1 microsecond, using starting conditions which consisted of fully extended protein chains. Disordered regions of FG nups which were simulated used the nup specific definitions defined in Yamada *et al* [3]. After equilibration, data was then taken from a 4 microsecond production run on which post simulation analysis was performed to determine the average radius of gyration of different domains as defined in Yamada *et al* [3]. Contact proximity between two given side-chain beads was calculated by counting the number of times side-chains were within 16 Angstroms of each other in the trajectory snapshots which were saved every 100 picoseconds during the 4 microseconds of production simulation, with contact number normalized by the number of snapshots to determine the contact probability.

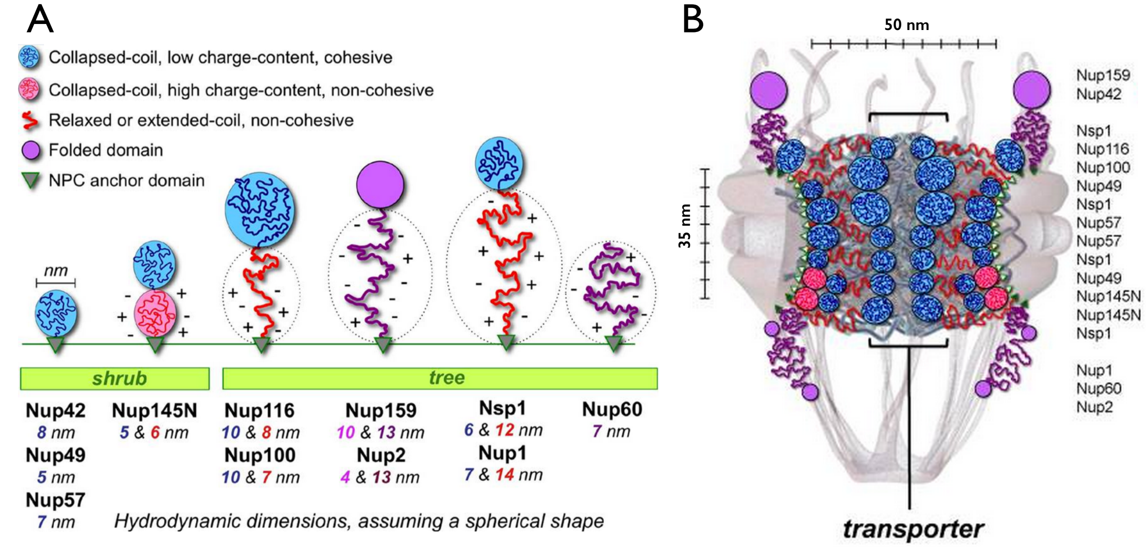

Figure 5.2: Biphasic FG nups and illustration of the Forest model. (A) Diagrams of the various FG nups in *S. cerevisiae* and their hydrodynamic radii. Blue or red lines depict either high (red) or low (blue) content of charged AAs along the disordered region of the FG nup. Purple lines represent nup domains with a net negative charge greater than 1. The small gray triangles represent the anchor domain of each FG nup. Single-block FG nups, termed "shrubs" in the Forest model, were categorized as consisting of a continuos collapsed FG domain adjacent to an anchor domain, consisting of Nup57, Nup49, and Nup42. Nup145N is shrub like, but not considered in this manuscript.

Figure 5.2: Di-block FG nups, or ”trees” in the terminology of the Forest model, were categorized as having a collapsed FG domain at their free ends separated from the anchor domain by a extended coil domain (B) A diagram of the Forest model NPC architecture. Orientation of the pore is such that top side faces the cytoplasm, while the bottom side faces the nucleus. FG nups are drawn to scale and positioned according to the relative location of their anchor domains along the z-axis of the NPC, as determined by immuno-localization [3]. These figures and research were originally published in Molecular and Cellular Proteomics. Yamada *et al.*, A Bimodal Distribution of Two Distinct Categories of Intrinsically Disordered Structures with Separate Functions in FG Nucleoporins. 2010; 9:2205-2224. ©The American Society for Biochemistry and Molecular Biology

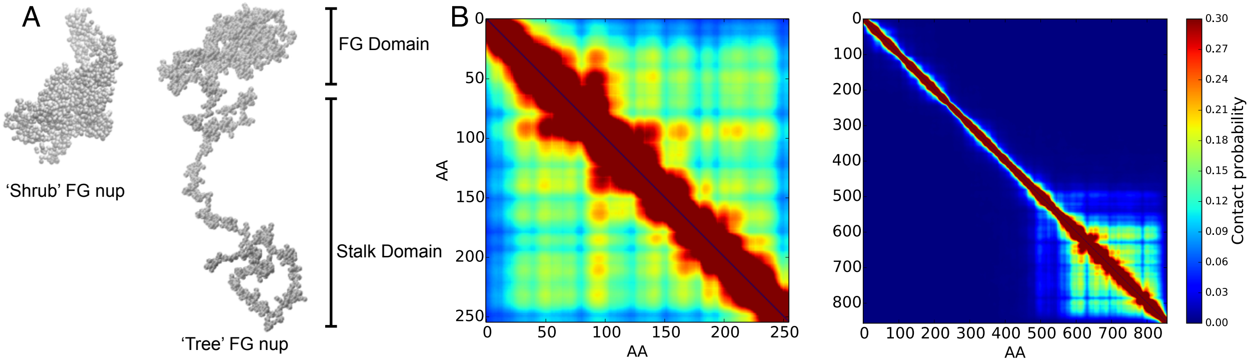

Figure 5.3: Coarse grained molecular modeling of single-block FG nups. (A) Left is a simulation snapshot of full length Nup57 showing the single polymer block FG domain. Right shows a snapshot of full length Nup1 demonstrating the di-block structure of FG and extended domain structures. (B) Left, the contact probability map for Nup57 show the time-averaged contacts between all pairs of amino acids. A monolithic block diagonal structure fills the entire contact map, implying that this FG nup is comprised of a single polymer block FG domain. Right, in contrast, the contact map for human Nup1 shows one block for the FG domain and diagonal contacts for the extended polymer domain, with both domains having low probability of cross contact.

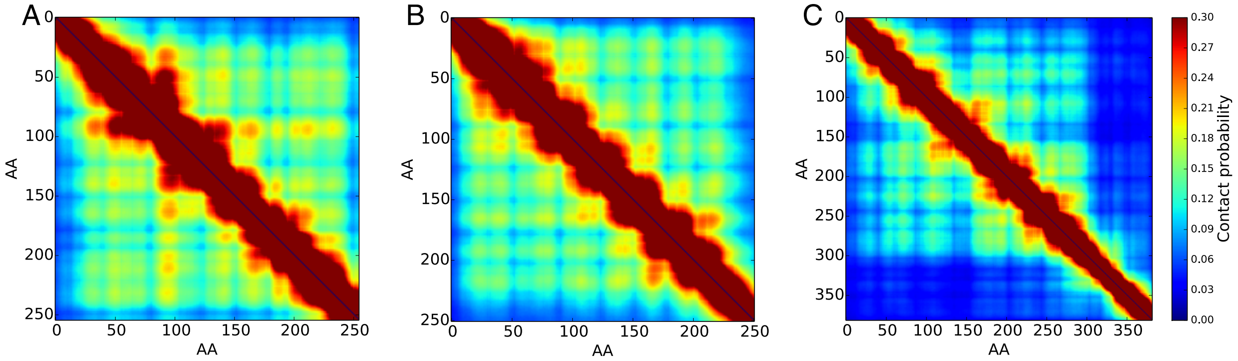

Figure 5.4: Contact maps for the different single-block FG nups in Baker's yeast from coarse grained molecular modeling. (A) Nup57's contact map (B) Nup49's contact map (C) Nup42's contact map. Contact probability maps show the time averaged contacts between all pairs of amino acids. A block diagonal structure in the contact maps represents a collapsed coil structure, which describes the single-block FG nups well, except for a small region of around 80 amino acids in length where Nup42 anchors to the pore wall. Amino acid residues shown are with respect to the disordered domains of the simulated FG nups, while full protein amino indexes can be determined by domain definitions in Yamada *et al* [3].

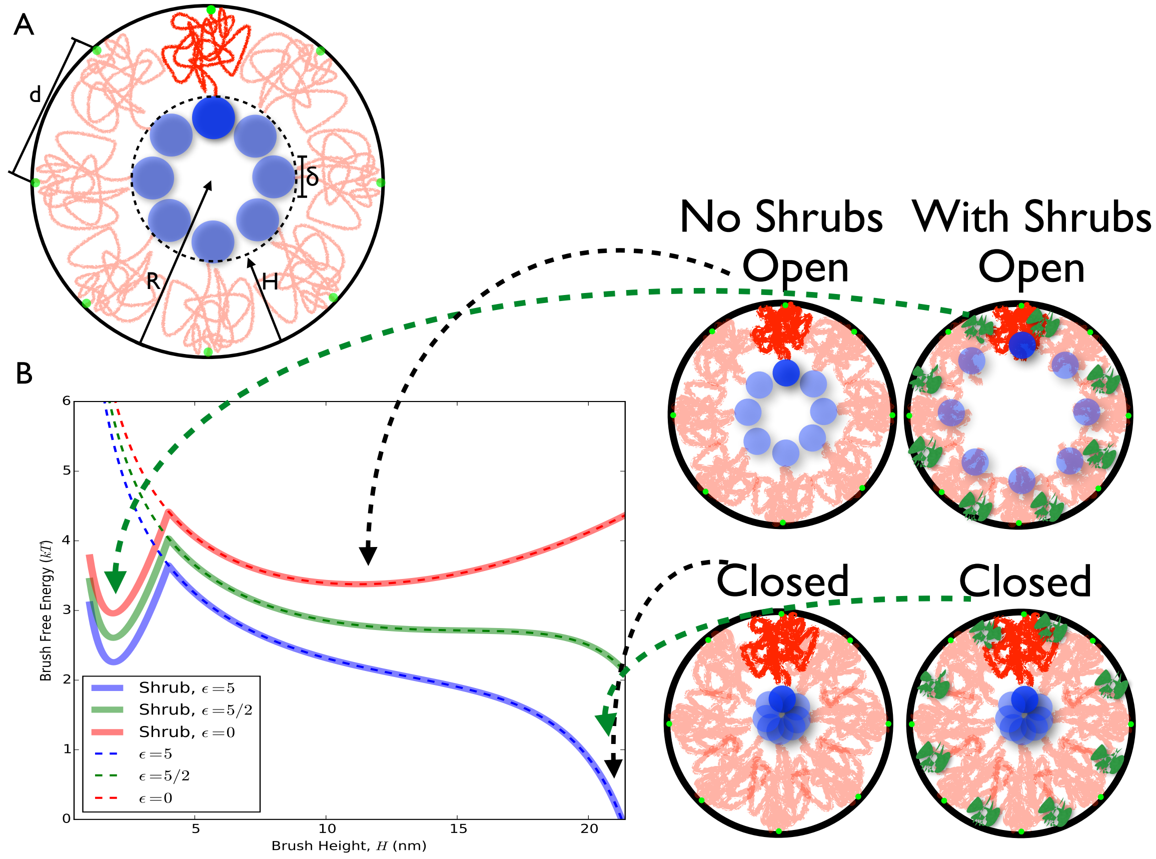

Figure 5.5: (A) Schematic of a polymer brush structure formed by di-block FG nups. Parameters H, height of the brush; R, radius of the pore; δ, diameter of the ’sticky tips’; and d, the average distance between anchor points. Green circles represent the locations at which FG nups are grafted to the pore. (B) Free energy of the Nsp1 brush, with and without single-block FG nups present on the pore wall, as a function of brush height for various values of the blob cohesive energy ($\epsilon k_B T$). Brush heigh can extend to a maximum of around 22 nm, which is the radius R of the modeled pore minus the size of the sticky tips. Di-block FG nup tip to single-block FG nup cohesion is fixed at $\epsilon_s = 6kT$. Right: Schematic diagram of the proposed Di-block Copolymer Brush Gate model at various minima of the brush free energy. When particular transport factors are present which are able to outcompete the inter-FG domain ”sticky tip” interactions, the brush is able to open up to a new free energy minimum that can accommodate the cargo. When interactions between sticky tips are able to recover into the several kT range, the pore is able to close with a free energy minimum at $H \sim R - \delta$. We have previously estimated the self interaction energy level of the Nsp1 sticky tip to be 4.7 kT [112], which also sets the energy scale for blob-blob interactions of ϵkT.

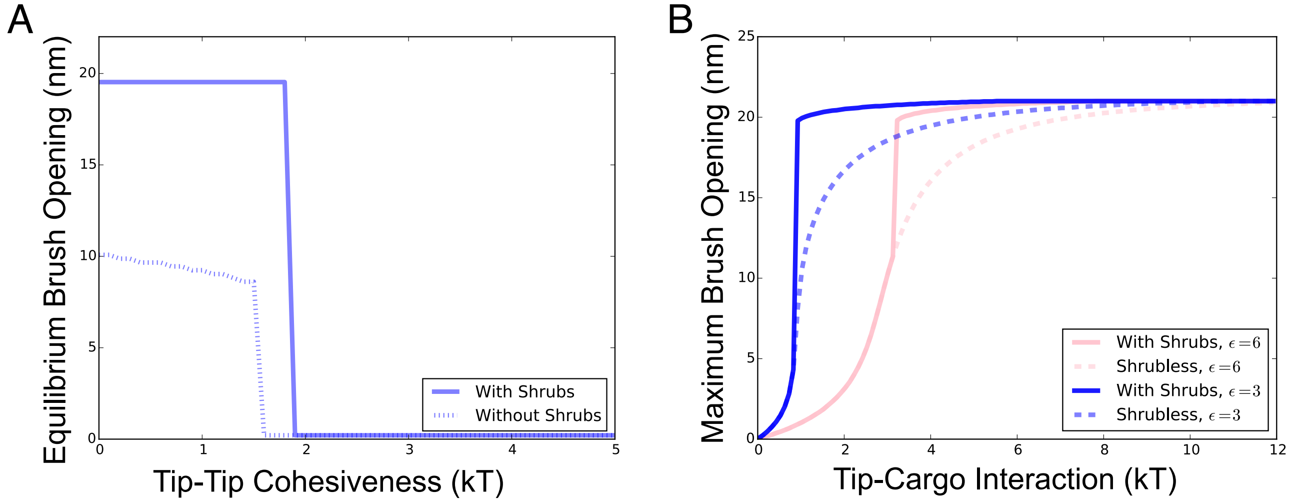

Figure 5.6: Overview of brush response to tip-tip and tip-transport factor interaction levels in the Di-block Copolymer Brush Gate (DCBG) model with single-block FG nups. (A) Changes in tip-tip cohesiveness result in equilibrium brush heights which are either significantly open or completely closed, with a sharp transition at ∼1.6 kT when single-block FG nups are not present and ∼1.9 kT when single-block FG nups are present. Tip to single-block FG nup cohesion is fixed at $\epsilon_s = 6kT$. (B) Tip-cargo interactions provide free energy by which work is done on the brush to reach a new equilibrium brush height. Brush height is plotted for different amounts of energy introduced into the brush system by tip-transport factor binding for different values of tip to top cohesion ϵ. Tip to single-block FG nup cohesion is fixed at $\epsilon_s = 6kT$.

Chapter 6

CONCLUSIONS

My first approach to understanding how the NPC regulates traffic focused on developing a novel form of bioinformatics for disordered proteins. This is especially challenging given the high mutation rates for disordered proteins and that functionality may not be strongly related to sequence. Based on the spatial clustering of physically relevant features such as binding motifs and charges within FG nups, I found a set of highly conserved spatial features in the sequence structure of individual FG nups, such as the separation, localization, and ordering of FG motifs and charged residues along the protein chain. These functionally conserved features provided a major insight into the particular biophysical mechanisms responsible for regulation of nucleocytoplasmic traffic in the NPC. From a bioinformatics perspective FG nups are diblock polymers, with one block hydrophobic and the other hydrophilic.

I further focused my research on trying to understand how the NPC regulates traffic by performing highly accurate but coarse grained molecular modeling of individual and aggregate FG nups to understand the dynamical and structural implications of the diblock structure of FG nups. Our results indicate that different regions or blocks of an individual FG nup can have distinctly different forms of disorder and that this property appears to be a conserved functional feature. A single FG nup has one block which is hydrophobic and collapsed while remaining dynamic and completely disordered, with another block that resembles an extended hydrophilic polymer. Furthermore, this block structure at the individual protein level was critical to the formation of a unique higher-order polymer brush architecture that can exist in distinct morphologies depending on the effective interaction energy between the hydrophobic domains of different FG nups. Because the interactions between hydrophobic domains may be modulated by certain forms of transport factors, our results indicate that cargo could induce transitions between different emergent brush morphologies which physically opens and closes the pore to regulate transport across the NPC. Furthermore, I found that adding a layer of ”shrub” FG nups along the NPC pore wall, as found in vivo, increases the size of the pore opening. This novel form of diblock polymer gated transport across membrane pores has wide biomimetic applicability.

My coarse grained simulations so far have yet to include cargo, with the molecular study of the dynamics of the nuclear pore as cargo translocate remaining unstudied. I therefore envision that future work regarding NPC transport studies should be focused on using very fast coarse grain molecular models such as dissipative particle dynamics (DPD) models to move beyond the equilibrium brush structure studies performed for the research in this dissertation.

Chapter 7

FUTURE WORK

7.1 Full pore DPD simulations

In preparation for future research with full pore simulations to understand the dynamics of transport I am building a molecular model of the NPC which is computationally highly efficient/parallel and allows for the simulation of FG nup aggregates over the long timescales for which transport events take place. This will allow for studying the dynamics of FG nup protein aggregates rather than just their equilibrium structure. Initially this model should be used to map the free energy landscape of the NPC as cargo of various size and hydrophobicity are inserted into the pore. Next steps would likely involve making the surface of the computational pore wall hydrophilic as found *in vivo*, which I predict will allow to pore to transport much larger cargo given that hydrophobic domains will likely stick to the pore wall when large cargo enter the NPC. This computational model of *in vivo* like NPCs, should be able to characterize cargo translocation times, cargo location during transport, and rejection rates for cargoes of various sizes and surface hydrophobicity. Finally, in order to design biomimetic *in vitro* pores which utilize the same brush gating mechanism as the NPC, I believe it would be best to simulate pores which have been filled with biphasic PEG-PPS to determine optimal PEG and PPS block lengths to regulate transport of cargos of a given size and surface chemistry. These results will likely lead in two different directions, one with a focus on the *in vivo* results and another which will focus on the design principles of biomimetic pores *in vitro*.

BIBLIOGRAPHY

[1] Dunker, A., Z. Obradovic, P. Romero, E. Garner, and C. Brown, 2000. Intrinsic Protein Disorder in Complete Genomes. *Proc. Genome Informatics.* 11:161–171.

[2] Dunker, A., C Brown, J. Lawson J,L. Iakoucheva LM, and Z. Obradovic, 2002. Intrinsic disorder and protein function. *Biochemistry.* 21:6573 6582.

[3] Yamada, J., J. L. Phillips, S. Patel, G. Goldfien, A. Calestagne-Morelli, H. Huang, R. Reza, J. Acheson, V. V. Krishnan, S. Newsam, et al., 2010. A bimodal distribution of two distinct categories of intrinsically disordered structures with separate functions in FG nucleoporins. *Molecular and Cellular Proteomics* 9:2205–2224.

[4] Alber, F., S. Dokudovskaya, L. M. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprapto, O. Karni-Schmidt, R. Williams, B. T. Chait, et al., 2007. The molecular architecture of the nuclear pore complex. *Nature* 450:695–701.

[5] Rout, M. P., J. D. Aitchison, A. Suprapto, K. Hjertaas, Y. Zhao, and B. T. Chait, 2000. The yeast nuclear pore complex composition, architecture, and transport mechanism. *The Journal of cell biology* 148:635–652.

[6] Yang, Q., M. P. Rout, and C. W. Akey, 1998. Three-Dimensional Architecture of the Isolated Yeast Nuclear Pore Complex: Functional and Evolutionary Implications. *Molecular Cell* 1:223–234.

[7] Macara, I. G., 2001. Transport into and out of the Nucleus. *Microbiology and Molecular Biology Reviews* 65:570–594.

[8] Denning, D. P., S. S. Patel, V. Uversky, A. L. Fink, and M. Rexach, 2003. Disorder in the nuclear pore complex: the FG repeat regions of nucleoporins are natively unfolded. *Proceedings of the National Academy of Sciences* 100:2450–2455.

[9] Bayliss, R., T. Littlewood, L. A. Strawn, S. R. Wente, and M. Stewart, 2002. GLFG and FxFG nucleoporins bind to overlapping sites on importin-β. *Journal of Biological Chemistry* 277:50597–50606.

[10] Pante, N., and M. Kann, 2002. Nuclear Pore Complex Is Able to Transport Macromolecules with Diameters of 39 nm. *Molecular Biology of The Cell* 13:425–434.

[11] Grünwald, D., and R. H. Singer, 2010. In vivo imaging of labelled endogenous [bgr]-actin mRNA during nucleocytoplasmic transport. *Nature* 467:604–607.

[12] Frey, S., R. P. Richter, and D. Görlich, 2006. FG-Rich Repeats of Nuclear Pore Proteins Form a Three-Dimensional Meshwork with Hydrogel-Like Properties. *Science* 314:815–817.

[13] Peters, R., 2005. Translocation Through the Nuclear Pore Complex: Selectivity and Speed by Reduction-of-Dimensionality. *Traffic* 6:421–427.

[14] Rout, M. P., J. D. Aitchison, M. O. Magnasco, and B. T. Chait, 2003. Virtual gating and nuclear transport: the hole picture. *Trends in Cell Biology* 13:622–628.

[15] Allen, N. P. C., L. Huang, A. Burlingame, and M. Rexach, 2001. Proteomic Analysis of Nucleoporin Interacting Proteins. *Journal of Biological Chemistry* 276:29268–29274.

[16] Allen, N. P. C., 2002. Deciphering Networks of Protein Interactions at the Nuclear Pore Complex. *Molecular and Cellular Proteomics* 1:930–946.

[17] Denning, D. P., and M. F. Rexach, 2006. Rapid Evolution Exposes the Boundaries of Domain Structure and Function in Natively Unfolded FG Nucleoporins. *Molecular and Cellular Proteomics* 6:272–282.

[18] Krishnan, V. V., E. Y. Lau, J. Yamada, D. P. Denning, S. S. Patel, M. E. Colvin, and M. F. Rexach, 2008. Intramolecular Cohesion of Coils Mediated by Phenylalanine–Glycine Motifs in the Natively Unfolded Domain of a Nucleoporin. *PLOS Computational Biology* 4.

[19] Zilman, A., S. DiTalia, B. T. Chait, M. P. Rout, and M. O. Magnasco, 2007. Efficiency, selectivity and robustness of the nuclear pore complex transport. *PLoS Comput Biology* 3.

[20] Zilman, A., S. Di Talia, T. Jovanovic-Talisman, B. T. Chait, M. P. Rout, and M. O. Magnasco, 2010. Enhancement of transport selectivity through nano-channels by non-specific competition. *PLoS computational biology* 6:e1000804.

[21] Bickel, T., and R. Bruinsma, 2002. The Nuclear Pore Complex Mystery and Anomalous Diffusion in Reversible Gels. *Biophysical Journal* 83:3079–3087.

[22] Osmanovic, D., J. Bailey, A. H. Harker, A. Fassati, B. W. Hoogenboom, and I. J. Ford, 2012. Bistable collective behavior of polymers tethered in a nanopore. *Physical Review E* 85:061917.

[23] Mincer, J. S., and S. M. Simon, 2011. Simulations of nuclear pore transport yield mechanistic insights and quantitative predictions. *Proceedings of the National Academy of Sciences* 108:E351–E358.

[24] Moussavi-Baygi, R., Y. Jamali, R. Karimi, and M. R. Mofrad, 2011. Brownian dynamics simulation of nucleocytoplasmic transport: a coarse-grained model for the functional state of the nuclear pore complex. *PLoS computational biology* 7:e1002049.

[25] Moussavi-Baygi, R., Y. Jamali, R. Karimi, and M. Mofrad, 2011. Biophysical coarse-grained modeling provides insights into transport through the nuclear pore complex. *Biophysical journal* 100:1410–1419.

[26] Tagliazucchi, M., O. Peleg, M. Kröger, Y. Rabin, and I. Szleifer, 2013. Effect of charge, hydrophobicity, and sequence of nucleoporins on the translocation of model particles through the nuclear pore complex. *Proceedings of the National Academy of Sciences* 110:3363–3368.

[27] Hills Jr, R. D., L. Lu, and G. A. Voth, 2010. Multiscale coarse-graining of the protein energy landscape. *PLoS computational biology* 6:e1000827.

[28] Lim, R. Y., U. Aebi, and B. Fahrenkrog, 2008. Towards reconciling structure and function in the nuclear pore complex. *Histochemistry and cell biology* 129:105–116.

[29] Rout, M., J. Aitchison, M. Magnasco, and B. Chait, 2003. Virtual gating and nuclear transport: the hole picture. *Trends in cell biology* 13: 622–628.

[30] Rout, M., J. Aitchison, A. Suprapto, K. Hjertaas, Y. Zhao, et al., 2000. The yeast nuclear pore complex composition, architecture, and transport mechanism. *The Journal of cell biology* 148: 635–652.

[31] Alber, F., S. Dokudovskaya, L. Veenhoff, W. Zhang, J. Kipper, et al., 2007. The molecular architecture of the nuclear pore complex. *Nature* 450: 695–701.

[32] Fahrenkrog, B., and U. Aebi, 2003. The nuclear pore complex: nucleocytoplasmic transport and beyond. *Nature Reviews Molecular Cell Biology* 4: 757–766.

[33] Fried, H., and U. Kutay, 2003. Nucleocytoplasmic transport: taking an inventory. *Cellular and molecular life sciences* 60: 1659–1688.

[34] Isgro, T., and K. Schulten, 2005. Binding dynamics of isolated nucleoporin repeat regions to importin-β. *Structure* 13: 1869–1879.

[35] Terry, L., and S. Wente, 2009. Flexible gates: dynamic topologies and functions for fg nucleoporins in nucleocytoplasmic transport. *Eukaryotic cell* 8: 1814–1827.

[36] DeGrasse, J., K. DuBois, D. Devos, T. Siegel, A. Sali, et al., 2009. Evidence for a shared nuclear pore complex architecture that is conserved from the last common eukaryotic ancestor. *Molecular & Cellular Proteomics* 8: 2119–2130.

[37] Atkinson, C., A. Mattheyses, M. Kampmann, and S. Simon, 2013. Fluorescence anisotropy reveals order and disorder of protein domains in the nuclear pore complex *Biophysical Journal* 104: 37–50.

[38] Hülsmann, B., A. Labokha, and D. Görlich, 2012. The permeability of reconstituted nuclear pores provides direct evidence for the selective phase model. *Cell* 150: 738–751.

[39] Cardarelli, F., L. Lanzano, and E. Gratton, 2012. Capturing directed molecular motion in the nuclear pore complex of live cells. *Proceedings of the National Academy of Sciences* 109: 9863–9868.

[40] Brown, C., A. Johnson, and G. Daughdrill, 2010. Comparing models of evolution for ordered and disordered proteins. *Molecular biology and evolution* 27: 609–621.

[41] McGinnis, S., and T. Madden, 2004. Blast: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research* 32: W20–W25.

[42] Henikoff, S., and J. Henikoff, 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* 89: 10915–10919.

[43] Dayhoff, M., and R. Schwartz, 1978. A model of evolutionary change in proteins. *Atlas of protein sequence and structure.* National Biomedical Research Foundation. pp. 345–352.

[44] Denning, D., and M. Rexach, 2007. Rapid evolution exposes the boundaries of domain structure and function in natively unfolded fg nucleoporins. *Molecular & Cellular Proteomics* 6: 272–282.

[45] Moesa, H., S. Wakabayashi, K. Nakai, and A. Patil, 2012. Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification. *Molecular BioSystems* 8(12.: 3262-3273.

[46] Vuzman, D., and Y. Levy, 2010. DNA search efficiency is modulated by charge composition and distribution in the intrinsically disordered tail. *Proceedings of the National Academy of Sciences* 107: 21004–21009.

[47] Uversky, V., and K. Dunker, 2012. Multiparametric analysis of intrinsically disordered proteins: looking at intrinsic disorder through compound eyes. *Analytical chemistry* 84: 2096–2104.

[48] Vucetic, S., C. Brown, K. Dunker, and Z. Obradovic, 2003. Flavors of protein disorder. *Proteins: Structure, Function, and Bioinformatics* 52: 573–584.

[49] Wu, C., R. Apweiler, A. Bairoch, D. Natale, W. Barker, et al., 2006. The universal protein resource, uniprot.: an expanding universe of protein information. *Nucleic acids research* 34: D187–D191.

[50] Xue, B., R. Dunbrack, R. Williams, A. Dunker, and V. Uversky, 2010. Pondr-fit: a meta-predictor of intrinsically disordered amino acids. *Biochimica et Biophysica Acta, BBA.-Proteins & Proteomics* 1804: 996–1010.

[51] Bohm, C., K. Railing, H. Kriegel, P. Kroger, 2004. Density connected clustering with local subspace preferences. *Fourth IEEE International Conference on Data Mining*. IEEE. pp. 27–34.

[52] Tasoulis, D., V. Plagianakos, M. Vrahatis, 2004. Unsupervised clustering of bioinformatics data. *European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems*, Eunite. pp. 47–53.

[53] Dunn, J., 1973. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics* 3: 32-59.

[54] Yamada, J., J. Phillips, S. Patel, G. Goldfien, and A. Calestagne-Morelli, et al., 2010. A bimodal distribution of two distinct categories of intrinsically disordered structures with separate functions in fg nucleoporins. *Molecular & Cellular Proteomics* 9: 2205–2224.

[55] Patel, S., B. Belmont, J. Sante, and M. Rexach, et al., 2007. Natively unfolded nucleoporins gate protein diffusion across the nuclear pore complex. *Cell* 129: 83–96.

[56] Bayliss, R., T. Littlewood, L. Strawn, S. Wente, and M. Stewart, 2002. Glfg and fxfg nucleoporins bind to overlapping sites on importin-β. *Journal of Biological Chemistry* 277: 50597–50606.

[57] El-Shami, M., D. Pontier, S. Lahmy, L. Braun, and C. Picart, et al., 2007. Reiterated wg/gw motifs form functionally and evolutionarily conserved argonaute-binding platforms in rnai-related components. *Genes & development* 21: 2539–2544.

[58] Nguyen Ba, A., B. Yeh, D. van Dyk, A. Davidson, and B. Andrews, et al., 2012. Proteome-wide discovery of evolutionary conserved sequences in disordered regions. *Science Signalling* 5: rs1.

[59] Colwell, L., M. Brenner, and K. Ribbeck, 2010. Charge as a selection criterion for translocation through the nuclear pore complex. *PLoS Comput Biol* 6: e1000747.

[60] Terry, L., S. Wente, 2007. Nuclear mrna export requires specific fg nucleoporins for translocation through the nuclear pore complex. *The Journal of cell biology* 178: 1121–1132.

[61] Lim, R., B. Fahrenkrog, J. Köser, K. Schwarz-Herion, and J. Deng, et al., 2007. Nanomechanical basis of selective gating by the nuclear pore complex. *Science* 318: 640–643.

[62] Mincer, J., and S. Simon, 2011. Simulations of nuclear pore transport yield mechanistic insights and quantitative predictions. *Proceedings of the National Academy of Sciences* 108: E351–E358.

[63] Akey, C., 1990. Visualization of transport-related configurations of the nuclear pore transporter. *Biophysical journal* 58: 341.

[64] Shaulov, L., and A. Harel, 2012. Improved visualization of vertebrate nuclear pore complexes by field emission scanning electron microscopy. *Structure* 20: 407–413.

[65] Liashkovich, I., A. Meyring, H. Oberleithner, and V. Shahin, 2012. Structural organization of the nuclear pore permeability barrier. *Journal of Controlled Release.*

[66] Grünwald, D., R. Singer, and M. Rout, 2011. Nuclear export dynamics of rna-protein complexes. *Nature* 475: 333–341.

[67] Ma, J., A. Goryaynov, A. Sarma, and W. Yang, 2012. Self-regulated viscous channel in the nuclear pore complex. *Proceedings of the National Academy of Sciences* 109: 7326–7331.

[68] Eliezer, D., 2009. Biophysical characterization of intrinsically disordered proteins. *Current opinion in structural biology* 19: 23–30.

[69] Tagliazucchi, M., O. Peleg, M. Kröger, Y. Rabin, and I. Szleifer, 2013. Effect of charge, hydrophobicity, and sequence of nucleoporins on the translocation of model particles through the nuclear pore complex. *Proceedings of the National Academy of Sciences* 110: 3363–3368.

[70] Grünwald, D., and R. Singer, 2011. Multiscale dynamics in nucleocytoplasmic transport. *Current opinion in cell biology* 24: 100–106.

[71] Lim, R., K. Ullman, and B. Fahrenkrog, et al., 2008. Biology and biophysics of the nuclear pore complex and its components. *International review of cell and molecular biology* 267: 299.

[72] Achtert, E., H. Kriegel, and A. Zimek, 2008. Elki: a software system for evaluation of subspace clustering algorithms. *Scientific and Statistical Database Management.* Springer, pp. 580–585.

[73] Murphy, L., A. Wallqvist, and R. Levy, 2000. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering* 13: 149–152.

[74] Betts, M., and R. Russell, 2003. Amino acid properties and consequences of substitutions. *Bioinformatics for geneticists* 317: 289.

[75] Ayton, G. S., W. G. Noid, and G. A. Voth, 2007. Systematic coarse graining of biomolecular and soft-matter systems. *MRS Bulletin* 32:929–934.

[76] Noid, W., J.-W. Chu, G. S. Ayton, and G. A. Voth, 2007. Multiscale coarse-graining and structural correlations: Connections to liquid-state theory. *The Journal of Physical Chemistry B* 111:4116–4127.

[77] Karanicolas, J., and C. L. Brooks, 2002. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Science* 11:2351–2361.

[78] Milles, S., and E. A. Lemke, 2011. Single molecule study of the intrinsically disordered FG-repeat nucleoporin 153. *Biophysical journal* 101:1710–1719.

[79] Tcherkasskaya, O., E. A. Davidson, and V. N. Uversky, 2003. Biophysical constraints for protein structure prediction. *Journal of proteome research* 2:37–42.

[80] Lim, R. Y., N.-P. Huang, J. Köser, J. Deng, K. A. Lau, K. Schwarz-Herion, B. Fahrenkrog, and U. Aebi, 2006. Flexible phenylalanine-glycine nucleoporins as entropic barriers to nucleocytoplasmic transport. *Proceedings of the National Academy of Sciences* 103:9512–9517.

[81] Ando, D., M. Colvin, M. Rexach, and A. Gopinathan, 2013. Physical Motif Clustering within Intrinsically Disordered Nucleoporin Sequences Reveals Universal Functional Features. *PloS one* 8:e73831.

[82] Rubinstein, M., and R. H. Colby, 2003. Polymer physics. OUP Oxford.

[83] Grünwald, D., and R. H. Singer, 2012. Multiscale dynamics in nucleocytoplasmic transport. *Current opinion in cell biology* 24:100–106.

[84] Fischer, H., I. Polikarpov, and A. F. Craievich, 2004. Average protein density is a molecular-weight-dependent function. *Protein Science* 13:2825–2828.

[85] Akey, C. W., 2010. The NPC-transporter, a ghost in the machine. *Structure* 18:1230–1232.

[86] Beck, M., F. Förster, M. Ecke, J. M. Plitzko, F. Melchior, G. Gerisch, W. Baumeister, and O. Medalia, 2004. Nuclear pore complex structure and dynamics revealed by cryoelectron tomography. *Science* 306:1387–1390.

[87] Colwell, L. J., M. P. Brenner, and K. Ribbeck, 2010. Charge as a selection criterion for translocation through the nuclear pore complex. *PLoS computational biology* 6:e1000747.

[88] Sevick, E., 1996. Shear swelling of polymer brushes grafted onto convex and concave surfaces. *Macromolecules* 29:6952–6958.

[89] Flory, P. J., 1971. Statistical mechanics of chain molecules. Macmillan.

[90] Isgro, T. A., and K. Schulten, 2005. Binding Dynamics of Isolated Nucleoporin Repeat Regions to Importin-beta. *Structure* 13:1869–1879.

[91] Liu, S. M., and M. Stewart, 2005. Structural Basis for the High-affinity Binding of Nucleoporin Nup1p to the Saccharomyces cerevisiae Importin-? Homologue, Kap95p. *Journal of Molecular Biology* 349:515–525.

[92] Strawn, L. A., T. Shen, N. Shulga, D. S. Goldfarb, and S. R. Wente, 2004. Minimal nuclear pore complexes define FG repeat domains essential for transport. *Nature cell biology* 6:197–206.

[93] Kim, J., A. Izadyar, N. Nioradze, and S. Amemiya, 2013. Nanoscale mechanism of molecular transport through the nuclear pore complex as studied by scanning electrochemical microscopy. *Journal of the American Chemical Society* 135:2321–2329.

[94] Ma, J., and W. Yang, 2010. Three-dimensional distribution of transient interactions in the nuclear pore complex obtained from single-molecule snapshots. *Proceedings of the National Academy of Sciences* 107:7305–7310.

[95] Ma, J., A. Goryaynov, A. Sarma, and W. Yang, 2012. Self-regulated viscous channel in the nuclear pore complex. *Proceedings of the National Academy of Sciences* 109:7326–7331.

[96] Yang, W., 2013. Distinct, but not completely separate spatial transport routes in the nuclear pore complex. *Nucleus* 4:0–1.

[97] Yang, W., 2011. Natively unfolded nucleoporins in nucleocytoplasmic transport: clustered or evenly distributed? *Nucleus* 2:10–16.

[98] Williamson, J. R., 2008. Cooperativity in macromolecular assembly. *Nature chemical biology* 4:458–465.

[99] Ribbeck, K., and D. Görlich, 2002. The permeability barrier of nuclear pore complexes appears to operate via hydrophobic exclusion. *The EMBO journal* 21:2664–2671.

[100] Shulga, N., and D. S. Goldfarb, 2003. Binding Dynamics of Structural Nucleoporins Govern Nuclear Pore Complex Permeability and May Mediate Channel Gating. *Molecular and Cellular Biology* 23:534–542.

[101] Jäkel, S., J.-M. Mingot, P. Schwarzmaier, E. Hartmann, and D. Görlich, 2002. Importins fulfil a dual function as nuclear import receptors and cytoplasmic chaperones for exposed basic domains. *The EMBO journal* 21:377–386.

[102] Martin, D. K., 2007. Nanobiotechnology of biomimetic membranes, volume 1. Springer.

[103] Bartsch, R. A., and J. D. Way, 1996. Chemical separations with liquid membranes: an overview. *In* ACS Symposium Series. ACS Publications, volume 642, 1–10.

[104] Peer, D., J. M. Karp, S. Hong, O. C. Farokhzad, R. Margalit, and R. Langer, 2007. Nanocarriers as an emerging platform for cancer therapy. *Nature nanotechnology* 2:751–760.

[105] Karlsson, M., M. Davidson, R. Karlsson, A. Karlsson, J. Bergenholtz, Z. Konkoli, A. Jesorka, T. Lobovkina, J. Hurtig, M. Voinova, et al., 2004. Biomimetic nanoscale reactors and networks. *Annu. Rev. Phys. Chem.* 55:613–649.

[106] Prokop, A., 2001. Bioartificial Organs in the Twenty-first Century. *Annals of the New York Academy of Sciences* 944:472–490.

[107] Photos, P. J., H. Bermudez, H. Aranda-Espinoza, J. Shillcock, and D. E. Discher, 2007. Nuclear pores and membrane holes: generic models for confined chains and entropic barriers in pore stabilization. *Soft Matter* 3:364–371.

[108] Jovanovic-Talisman, T., J. Tetenbaum-Novatt, A. S. McKenney, A. Zilman, R. Peters, M. P. Rout, and B. T. Chait, 2008. Artificial nanopores that mimic the transport selectivity of the nuclear pore complex. *Nature* 457:1023–1027.

[109] Plimpton, S., 1995. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics* 117:1–19.

[110] Xue, B., R. L. Dunbrack, R. W. Williams, A. K. Dunker, and V. N. Uversky, 2010. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1804:996–1010.

[111] Rubinstein, M., and R. H. Colby, 2003. Polymer physics. *OUP Oxford.*

[112] Ando, D., R. Zandi, Y. Kim, M. Colvin, M. Rexach, and A. Gopinathan, 2014. Nuclear Pore Complex Protein Sequences Determine Overall Copolymer Brush Structure and Function. *Biophysical Journal* 9:1997 – 2007.

[113] Ghavami, A., L. Veenhoff , E. van der Giessen, and P. Onck, 2014. Probing the disordered domain of the nuclear pore complex through coarse-grained molecular dynamics simulations. *Biophysical journal* 107:1393–1402

Appendix A

ABBEVIATIONS & NOMENCLATURE

A.1 Abbreviations

FG = Phenylalanine glycine
FG nups = FG nucleoporins
NPC = Nuclear Pore Complex
IDP = Intrinsically Disordered Protein
CG = Coarse grain
DPD = Dissipative particle dynamics
DCBG = Diblock Copolymer Brush Gate

A.2 Nomenclature

H = Brush height [nm]
R = Radius of the pore [nm]
R_s = Radial location of shrub FG nups [nm]
δ = Diameter of the sticky tip blobs [nm]
δ_s = Diameter of the shrub blobs [nm]
ϵ = Cohesive energy between tip blobs [kT]
ϵ_s = Cohesive energy between tip blobs and shrub FG nups [kT]
d = Average distance between anchor points [nm]
a = Length of monomers in the chains [nm]
T = Temperature [K]