Rapid Evolution Exposes the Boundaries of Domain Structure and Function in Natively Unfolded FG Nucleoporins*

Daniel P. Denning‡ and Michael F. Rexach§

Nucleoporins with phenylalanine-glycine repeats (FG Nups) function at the nuclear pore complex (NPC) to facilitate nucleocytoplasmic transport. In Saccharomyces cerevisiae, each FG Nup contains a large natively unfolded domain that is punctuated by FG repeats. These FG repeats are surrounded by hydrophilic amino acids (AAs) common to disordered protein domains. Here we show that the FG domain of Nups from human, fly, worm, and other yeast species is also enriched in these disorder-associated AAs, indicating that structural disorder is a conserved feature of FG Nups and likely serves an important role in NPC function. Despite the conservation of AA composition, FG Nup sequences from different species show extensive divergence. A comparison of the AA substitution rates of proteins with syntenic orthologs in four Saccharomyces species revealed that FG Nups have evolved rapidly (particularly in their FG domains) and most substitutions occurring in sequences between FG repeats. The rapid evolution of FG Nups is poorly explained by parameters known to influence AA substitution rate, such as protein expression level, interactivity, and essentiality; instead their rapid evolution may reflect an intrinsic permissiveness of natively unfolded structures to AA substitutions. The overall lack of AA sequence conservation in FG Nups is sharply contrasted by discrete stretches of conserved sequences. These conserved sequences highlight known karyopherin and nucleoporin binding sites as well as other uncharacterized sites that may have important structural and functional properties.


In eukaryotes, nuclear pore complexes (NPCs)1 regulate the movement of cellular material across the nuclear envelope by functioning as a permeability barrier and as transport machine (1). Small molecules diffuse through NPCs, but large proteins and RNAs (>40 kDa) are excluded unless they contain localization sequences that permit translocation. These targeting signals are recognized by karyopherins (Kaps), mobile receptors that interact with the NPC to facilitate nucleocytoplasmic transport (2). Multiple copies of 30 nucleoporin (Nup) proteins comprise each NPC (3, 4), and approximately half of these Nups are classified as FG Nups due to their content of phenylalanine-glycine (FG) motifs. In Saccharomyces cerevisiae, each FG Nup contains a large domain (150–700 AA in length) composed of FG repeats spaced 10–20 AAs apart. These FG domains function as docking sites for Kaps (5), which bind to phenylalanines in the FG motif (6, 7).

The FG domains of S. cerevisiae Nups are natively unfolded (8, 9). Such “natively unfolded,” “intrinsically unstructured,” and “disordered” proteins or protein domains lack stable secondary structure and behave as flexible filaments (10, 11). Despite their structural disorder, these domains are often essential for protein-protein and protein-nucleic acid interactions (11). Although the role of disordered structure in FG Nup function is unclear, two hypotheses have been proposed. First the disordered structures in Nups may facilitate rapid translocation of Kap-cargo complexes through the NPC by capturing and releasing Kaps with fast association and dissociation rates. Disordered proteins can exhibit unusually rapid interaction dynamics with binding partners due to a lack of steric limitations (11). In nuclear transport, fast interactions between Kaps and Nups may be necessary for the rapid flux of Kap-cargo complexes through NPCs (12). Thus, the disordered structures of FG Nups might be optimized for highly specific, yet transient interactions with a variety of transport factors. A second hypothesis proposes that the NPC permeability barrier is a meshwork of disordered FG Nup filaments that are interconnected by weak hydrophobic interactions between FG motifs (12). In principle, such a barrier could permit small particles to pass through the interfilament space yet exclude larger molecules from entering the NPC. Large Kap-cargo complexes could gain access by interacting with multiple FG Nups.

If structural disorder in FG Nups serves a critical role in NPC function and architecture, then this feature should be preserved throughout Eukaryotae. In the present study we analyzed the AA composition of FG Nups from evolutionarily distant eukaryotes and found evidence of structural disorder in nearly all FG Nups examined. We also noted that FG Nups have evolved rapidly (particularly in their FG domains) and

1 The abbreviations used are: NPC, nuclear pore complex; Nup, nucleoporin; FG Nup, nucleoporin with phenylalanine-glycine repeats; FG domain, FG repeat region of a Nup; Kap, karyopherin; AA, amino acid; RRM, RNA recognition motif; AACB, AA composition bias; CAI, codon adaptation index.
evaluated the contribution of several parameters to their high amino acid substitution rates. Lastly we identified discrete regions of AA sequence conservation in the FG domains that coincide with known Kap and Nup binding sites and identified clusters of conserved AAs in the regions flanking FG domains that correlate with known NPC anchoring domains and with known molecular interaction sites.

**EXPERIMENTAL PROCEDURES**

**Nucleotide and AA Sequences—**The nucleotide and AA sequences of *S. cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, and *Saccharomyces bayanus* nucleoporins were acquired from the *Saccharomyces* Genome Database (yeastgenome.org) (13). The AA sequences of human FG Nups (4) were acquired from the Swiss-Prot (us.expasy.ch) and NCBI (ncbi.nlm.nih.gov) databases. The *Caenorhabditis elegans* FG Nups were identified previously by homology (14). Sequences for *C. elegans* FG Nups and their corresponding Drosophila melanogaster homologs were acquired from Wormbase (wormbase.org). The sequences of *Schizosaccharomyces pombe* FG Nups were obtained by searching the *S. pombe* Gene Data Bank (genedb.org) for FG-containing Nups.

To calculate the fraction of perfectly conserved AAs in yeast FG domains, the FG domains of Nups in *S. bayanus*, *S. mikatae*, and *S. paradoxus* were aligned against the *S. cerevisiae* FG domain. All of the FG domains used in this study are listed in Supplemental Table S3. The FG domains were defined as the largest contiguous sequence of FG repeats separated by less than 100 AAs and included 10 additional residues flanking the first and last FG motif. Insertions greater than 10 AAs within the FG domains were excluded, such as insertions in the *S. paradoxus* Nup159 ortholog. Deletions in the FG domains of Nup2 and Nup145 orthologs from *S. paradoxus* and *S. bayanus*, respectively, precluded the use of those regions; however, the remaining FG domain sequences available were used. The FG domain of Nsp1 was not included because Nsp1 orthologs were not initially available. FG domains were not defined for Nup53 and Nup59 or their homologs (spomNup40, NPP-19, dmNup35, and hsNup35) because their FG motifs are not clustered into obvious domains. AA composition, codon adaptation indices (CAIs), and Protein Data Bank homologs for *S. cerevisiae* proteins were taken from *Saccharomyces* Genome Database datasets. Protein Data Bank homologs were considered highly significant if a Smith-Waterman sequence analysis yielded a p value <10^{-10}.

**Evolution Rate Estimations—**Non-synonymous (dN) and synonymous (dS) substitution rates (the frequency of non-synonymous substitutions per non-synonymous site and the frequency of synonymous substitutions per synonymous site, respectively) for 2,956 proteins were calculated by Wall et al. (15) using gene sequences from *S. cerevisiae*, *S. bayanus*, *S. mikatae*, and *S. paradoxus* (13). This data set includes only genes with syntenic orthologs in all four yeast species with >80% coding sequence alignment and excludes ortholog sets with frame-shifts caused by insertions or deletions.

**AA Sequence Alignments—**The AA sequence alignments of Nups and Kaps from the four *Saccharomyces* species were generated using the Synteny Viewer program in the “Homology and Comparisons” section of the *Saccharomyces* Genome Database website (yeastgenome.org). The complete NSP1 sequences were identified after removing an intron and were aligned manually.

**Protein Translation Rates, Interactivity, and Dispensability—**Translation rates for 2,700 *S. cerevisiae* proteins were acquired from Arava et al. (16); the rates were estimated from the number of ribosomes that co-purify with mRNA transcripts. Protein interaction data were obtained from a compilation of large scale protein interaction screens, including yeast two-hybrid screens and biochemical tandem affinity purifications (17). For our analyses, protein-protein interactions were included only if they were identified independently in two or more of the large scale interaction screens.

*S. cerevisiae* proteins were identified as “essential” or “slow growth” based on results from the systematic deletion of *S. cerevisiae* genes (18, 19). Yeast lacking essential proteins are inviable on rich medium at 30 °C, whereas yeast lacking slow growth proteins grow at a reduced rate compared with wild-type yeast (19).

**RESULTS**

**The AA Composition of Most FG Nups Is Consistent with Structural Disorder—**We showed previously that the large FG domains of *S. cerevisiae* Nups are natively unfolded and are enriched in AAs that are common in disordered protein domains (8, 9). Generally disordered polypeptides are enriched in charged and polar AAs (Ala, Arg, Gin, Glu, Gly, Lys, Pro, and Ser) and are depleted of hydrophobic AAs (Asn, Cys, Ile, Leu, Phe, Trp, Tyr, and Val) (11); these two groups of AAs have been referred to as “disorder”-associated and “order”-associated, respectively. To examine whether structural disorder is a conserved feature of FG Nups in other eukaryotes, we determined the disorder- and order-associated AA content of FG Nups from *Homo sapiens*, *D. melanogaster*, *C. elegans*, *S. pombe*, *S. paradoxus*, *S. mikatae*, and *S. bayanus*. We found that nearly all FG Nups in these organisms exhibit a similar enrichment of disorder-associated AAs (Table I and Supplemental Table S3) as observed previously with the *S. cerevisiae* FG Nups (9). In particular, the FG domains show a higher concentration of disorder-associated AAs compared with the full-length FG Nup sequences, indicating that the FG domains are more likely to adopt disordered structure than non-FG domains. Thus, it appears that structural disorder of FG Nups is evolutionarily conserved.

**The Rapid Evolution of FG Nups—**Despite the abundance of FG repeats, similar AA compositions, and highly analogous functions in nuclear transport, the primary structures of FG Nups are poorly conserved over large evolutionary distances, making it difficult to identify orthologous genes in distantly related species (data not shown). In contrast, Kaps are well conserved between species; for example, the yeast Kap95 and human importin-β share 52% similarity over 99% of their sequence. Thus, it seems that the FG Nups have diverged significantly, whereas their Kap binding partners have not.

The poor sequence conservation between FG Nups throughout Eukaryotae suggests that FG Nups have evolved rapidly. To assess this, we compared the evolution rates of *Saccharomyces* FG Nups to those of thousands of other *Saccharomyces* proteins using AA substitution rates (dN) determined by Wall et al. (15). These rates were calculated using sequence alignments of syntenic orthologs from four *Saccharomyces* species: *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. bayanus*. The analysis showed that FG Nups have evolved on average 2 times faster than the mean yeast protein, 2 times faster than non-FG Nups, 3 times faster than Kaps, and 5 times faster than a set of 105 structured proteins (Fig. 1A).
Rapid protein evolution rates can be due either to weak purifying (negative) selection or to positive (adaptive) selection for advantageous amino acid substitutions. To distinguish between these possibilities, we examined the ratio of non-synonymous and synonymous substitution rates (dN/dS) for each yeast FG Nup: a dN/dS ratio <1 is evidence for purifying selection, a ratio =1 indicates neutral evolution, and a ratio >1 is evidence for positive selection. We examined the dN and dS values calculated by Wall et al. (15) for each FG Nup and observed that all of the FG Nups have dN/dS ratios less than or equal to 0.3 with a mean value of 0.2 (data not shown). The small ratios indicate that the FG Nups on average are under purifying selection, not positive selection. However, this analysis assumes a constant dN/dS ratio for the entire length of each FG Nup and might therefore mask individual codons subject to positive selection. A recent study examined 4,133 aligned genes from S. cerevisiae and S. paradoxus to identify genes with codons under positive selection (20). By allowing dN and dS to vary within each gene, this analysis identified 126 proteins, including Nup42 but no other Nups, with at least one codon under positive selection. We conclude that the vast majority of codons within the FG Nups are not under positive selection. Instead their overall rapid evolution rates likely reflect weak constraints on amino acid usage at most sites within their sequences.

A comparison of the non-identical AA sites in alignments of FG Nups from the four Saccharomyces species showed that divergent sites occur more frequently within FG domains than in non-FG domains (Fig. 1B). Indeed the low AA sequence conservation in the FG domains of Nups contrasts with the high level of conservation in Kaps and non-FG Nups (Figs. 1A and 3). Notably AA substitutions were highest among the FG domains of Nups located in peripheral NPC structures (e.g. Nup159, Nup42, Nup60, Nup1, and Nup2, which form the cytoplasmic fibers and nuclear basket) and were lowest among FG domains of Nups located in the central transport conduit (e.g. Nup116, Nup100, Nup57, and Nup49) (Fig. 1B).

An alignment of three major types of FG motifs (SAFGX-PSFG, GLFG, and FXFG) in Nups from the four Saccharomyces species showed that the majority of the divergent sites occur in the spacer regions that separate the FG motifs (Fig. 2). In these four species, a low percentage (~35%) of amino acids are perfectly conserved in sequences flanking SAFGX-PSFG motifs at positions −10 to −4 and +4 to +10 in relation to the phenylalanine (top panel). Similarly ~50 and ~35% of those positions are conserved in sequences flanking the GLFG and FXFG motifs, respectively (middle panels). In contrast, the Phe and Gly residues in FG motifs are under strong purifying selection as ~90% of the phenylalanines and ~80% of the glycines are conserved (bottom panel). This is consistent with crystal structures of the Kap-Nup interaction that show that the phenylalanine in FG motifs is the key binding determinant for Kaps (6). The conserved glycine residue also seems to be important for most FG motifs except for the FXFG motifs in Nup1 and Nup60 that lack or show poor conservation of the glycine residue (data not shown). Last the AAs immediately preceding the FG motifs at positions −3, −2, and −1 are also conserved (all panels). These positions classify FG motifs into subtypes, such as GLFG and FXFG, and may be important in determining the strength and specificity of interactions with Kaps.

"Islands" of AA Sequence Conservation in the FG Domains of Nups—The overall lack of AA conservation in the FG domains of Nups is contrasted by the presence of small, discrete stretches of conserved AAs with perfect sequence identity in all four yeast FG Nup orthologs. These islands of conserved AA sequences are present throughout the FG domains and are highlighted in yellow boxes in Fig. 3 and Supplemental Fig. S6. Typically the islands are 6–11 AA in length and center on a single conserved FG motif, although a minority of the islands include conserved AA sequences in between two FG motifs (Supplemental Fig. S7).

The length of most conserved islands in the FG domains (~8 residues long) is similar to the length of Nup1 sequences that contact Kap95 in co-crystals (6, 7, 21), suggesting that each of the conserved islands within the FG domains might function as a contact site for Kaps. Kap95 binds to the C terminus of Nup1 with high affinity (22) by making contact with...
two stretches of 13 and 6 AAs (AAs 975–987 and 1004–1009) (21). These contact sites are clearly delineated by conserved Boxes 22 and 23 in the otherwise highly substituted FG domain of Nup1 (Fig. 3). Box 22 contains 9 conserved AAs between two conserved Phe residues, and Box 23 has 4 conserved AAs flanking an FG motif. These results suggest that pairs of Phe residues linked by conserved AAs may represent high affinity binding sites for Kaps. Similarly the nine SAFG*PSFG motifs in Nup159 and Nup42 are also highly conserved (Fig. 2); these motifs contain two closely positioned FG repeats and provide high affinity binding sites for the exportin Crm1.2

Fig. 1. The FG Nups in budding yeast have evolved rapidly. A, the mean AA substitution rates for FG Nups, non-FG Nups, Kaps, a set of 105 yeast proteins with highly significant homologs in the Protein Data Bank of NMR and crystal structures, and all yeast proteins were derived from alignments of syntenic orthologs from S. cerevisiae, S. paradoxus, S. mikatae, and S. bayanus (analysis of variance, \( p = 2.2 \times 10^{-5} \)). Error bars represent 95% confidence intervals for each mean. B, the fraction of conserved AAs in the FG domains and non-FG domains calculated from alignments of the FG Nups of four Saccharomyces species.

It is also interesting to note that the centrally located GLFG Nups (Nup49, Nup57, Nup100, Nup116, and nNup145) contain stretches of conserved AAs at their N termini with two or more FG motifs (Supplemental Fig. S7); these stretches might be high affinity binding sites for Kaps. Additional examples of two FG motifs linked by conserved AAs can be found in other FG domains of Nups, most notably in Nup100 and Nup116 (Supplemental Fig. S7). For example, Box 6 in the FG domain of Nup116 (Fig. 3) contains two FG repeats and two additional phenylalanines and binds with high affinity to the non-FG Nup Gle2 (23). Finally the high affinity binding site for Kap121 on Nup53 includes Box 10 (Supplemental Fig. S6), which contains conserved AAs between two phenylalanines (24).

We also noted less common but potentially important FG motifs in the Nups of the four Saccharomyces species (Supplemental Fig. S8). These include the SLFG motif (in Nup100, Nup116, Nup49, nNup145, and Nsp1), the SPFG motif (in Nup42, Nup100, and Nup116), the SFG motif (in Nsp1, Nup49, and Nup1), the NXFG motif (in Nup49, Nup57, Nup116, Nup100, and Nsp1), the WLF motif (in Nup53 and Nup59), and the FXXFG motif (in Nup159, Nup42, Nup116, Nup100, Nup2, and Nup59). Two additional compound FG motifs, sequences with closely spaced FG repeats, were also noted: the triple FG motif in Nup159 and Nup42 and the quadruple Phe motif in Nup53 and Nup59 (Supplemental Fig. S8). In principle, each FG motif might mediate a specific set of Kap interactions and may specify the strength of these interactions as observed previously for entire FG domains (5).

\[ ^2 \text{L. Liu, C. Deng, and M. Rexach, personal communication.} \]
quadruple Phe motif in Nup53 and Nup59 may be part of a predicted RNA recognition motif (RRM) fold and may function in homodimer formation (25). Other stretches of conserved sequence without FG motifs also exist in the various FG domains (Fig. 3 and Supplemental Fig. S6); these sequences may represent non-canonical binding sites for Kaps or may delineate binding sites for Nups or other nuclear transport factors.

**AA Sequence Conservation in Non-FG Domains of Nups**—In comparison with the high AA substitution rates observed for the FG domains of Nups, most of the non-FG domains have retained longer stretches of conserved AAs. The conserved sequence stretches are often contiguous and can be grouped into larger “clusters” of AAs, which are denoted by blue boxes in Fig. 3 and Supplemental Fig. S6. Many of these clusters correspond to domains with previously characterized functional or structural features (Table II and references therein). For example, the conserved Cluster I in Nsp1 and Nup116 clearly defines their respective NPC anchoring domains (26, 27), which are predicted to form coiled coil structures in Nsp1 (27) and a β-sheet structure in Nup116 (28). Cluster I in the N terminus of Nup159 demarcates a β-propeller structure that interacts with the DEAD box helicase Dbp5 (29), and Cluster II in Nup2 coincides with its C-terminal Ran binding domain (30). Additional examples of conserved non-FG sequences with characterized functions or structures are shown in Table II, including the binding site for importin-α (Kap60) at the N terminus of Nup2 (31, 32) and the predicted RRM domains of Nup53 and Nup59 (25). The correlation between conserved AA sequences in FG Nups and domains with previously characterized structure or function implies that other NPC anchoring domains or functionally important domains can be accurately predicted from the alignments shown in Fig. 3 and Supplemental Fig. S6. Based on that premise, we list all characterized and predicted domains of FG Nups in Table II and provide a refinement of their AA boundaries based on the extent of AA conservation.

**The Influence of Protein Expression, Interactivity, and Essentiality on FG Nup Evolution**—To explain the overall fast evolution rate of FG Nups, we examined each for characteristics that correlate with high AA substitution rates, such as low expression level, non-essential function, and low numbers of interacting proteins (33–35). These characteristics contrast with those of highly conserved proteins, which include abundant proteins that must retain codons recognized by abundant tRNAs (36, 37), essential proteins that are required for viability (38, 39), and proteins with many binding partners, which have a high fraction of AA residues under purifying selection (40, 41).

Using available experimental and bioinformatics data, we examined the substitution rate (dN), CAI, translation rate, and protein interactivity (binding partners per protein) of each FG Nup in relation to the S. cerevisiae proteome. As expected, the high percentile rank for FG Nup substitution rates shows that they are among the fastest evolving proteins in the yeast proteome (Fig. 4). In contrast, the FG Nups exhibit a wide range of CAI values, translation rates, and numbers of binding partners. Therefore, none of these parameters appears to explain the high substitution rates of FG Nups, implying that a different property is responsible for their rapid evolution.

**A Relationship between Structural Disorder and High Evolution Rates**—It is generally assumed that natively unfolded protein domains (such as the FG domains of Nups) are inherently more permissive to AA substitutions than domains with ordered, folded structures. Unfolded protein domains may be subject to less stringent sequence constraints as charged or polar AAs that promote interactions with the aqueous environment may substitute freely without compromising the overall disorder and flexibility of the FG domains. Indeed a study of 26 proteins with disordered domains reported high AA substitution rates in 19 of the 26 (42), suggesting that disordered structures and high AA substitution rates may be generally correlated. To further explore this relationship, we compared the overall structural properties and AA substitution rates of each Nup. To estimate structural content, we adapted the AA classifications determined by Dunker et al. (11) to generate a numerical value that reflects the enrichment of disorder- or order-associated AAs in each Nup. This AA composition bias (AACB) value is expressed as the percentage of disorder-associated AAs in a protein minus the percentage of order-associated AAs. For the S. cerevisiae Nups (FG and non-FG), we observed a strong correlation between AACB value and AA substitution rate (dN) (linear regression: dN = 0.0065(AACB) + 0.168, R^2 = 0.369) (Fig. 5A). As expected, Nups with the highest content of order-associated AAs (mostly the non-FG Nups) have fewer AA substitutions than Nups with the highest content of disorder-associated AAs (mostly the FG Nups). In contrast, a set of 105 yeast proteins with folded structures yielded a weak correlation between substitution rate and AACB value (Fig. 5B; linear

---

**Fig. 3. AA sequence alignment of Nups and a Kap from four Saccharomyces species.** Identical AAs in all four species are depicted in black bold font, whereas AAs substituted in at least one yeast are depicted in gray. Contiguous conserved sequences of ≥4 identical AA are boxed in yellow and numbered at top left for reference. Yellow boxes are extended in cases where a single non-identical AA separates two or more conserved AAs. Conserved Phe residues are highlighted in red bold font, and non-conserved Phe residues are orange. Clusters of conserved boxes are grouped in blue for non-FG domains and in red for FG domains. Extended FG domains (demarcated by the thin red line boxes) contain the FG domain plus flanking regions that are also highly substituted. Clusters in non-FG domains are separated when ≥10 contiguous, non-conserved AAs lie in between them. Nup53 and Nup59 have no clear FG domain; thus, only clusters were assigned. Dotted blue boxes mark the boundaries of previously characterized domains.
### Table II

A survey of known structures, functions, and interaction domains for each FG Nup with refined domain boundaries based on AA conservation.

Known structures and functions are in regular text; previously predicted structures and functions are in *italics*; new predictions of structure and function based on AACB value and AA sequence conservation are in **bold italics**. FG domain plus regions include the FG domain plus flanking, highly substituted sequences. hs RNA, heat shock RNA.

<table>
<thead>
<tr>
<th>Nup domains defined by AA conservation</th>
<th>Domain AAs</th>
<th>Percent AA conservation</th>
<th>AA composition (percent disorder/order)</th>
<th>AACB value</th>
<th>Actual or predicted structure (AAs)</th>
<th>Actual or predicted function (AAs)</th>
<th>Refs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nup159</td>
<td>Cluster I  1–386</td>
<td>68</td>
<td>45/41</td>
<td>4</td>
<td>β-Propeller (2–387)</td>
<td>Dbp5 binding (2–387)</td>
<td>29</td>
</tr>
<tr>
<td>FG domain plus</td>
<td>387–1071</td>
<td>20</td>
<td>62/22</td>
<td>40</td>
<td>Natively unfolded (441–881)</td>
<td>Kap binding (441–881)</td>
<td>9, 46</td>
</tr>
<tr>
<td>Cluster II 1072–1266 65 48/35 13</td>
<td>Structured</td>
<td>Coiled coil (1281–1316)</td>
<td>Molecular interaction Structural bridge or spacer</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Clusters II and III 1267–1331</td>
<td>25</td>
<td>38/43</td>
<td>−5</td>
<td>Coiled coil (1392–1412)</td>
<td>NPC anchor, Nup82 binding (1223–1460)</td>
<td>48–50</td>
<td></td>
</tr>
<tr>
<td>Cluster III 1332–1457 67 49/36 13</td>
<td>Coiled coil</td>
<td>13</td>
<td>Natively unfolded</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nup42</td>
<td>FG domain plus 1–382</td>
<td>22</td>
<td>56/29</td>
<td>27</td>
<td>Natively unfolded</td>
<td>Kap binding (1–430)</td>
<td>5</td>
</tr>
<tr>
<td>Cluster I 383–430 79 54/37 17</td>
<td>Structured</td>
<td>hs RNA export; NPC anchor (365–430)</td>
<td>51, 52</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nup49</td>
<td>FG domain plus 1–251</td>
<td>60</td>
<td>59/30</td>
<td>29</td>
<td>Natively unfolded</td>
<td>Kap binding (1–472)</td>
<td>5</td>
</tr>
<tr>
<td>Cluster I 252–458 90 45/37 8</td>
<td>Coiled coil (350–400)</td>
<td>8</td>
<td>Natively unfolded</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nup57</td>
<td>FG domain plus 1–255</td>
<td>61</td>
<td>58/26</td>
<td>32</td>
<td>Natively unfolded</td>
<td>Kap binding (1–541)</td>
<td>5</td>
</tr>
<tr>
<td>Cluster I 256–530 89 53/34 19</td>
<td>Coiled coil (350–425)</td>
<td>19</td>
<td>NPC anchor</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nup100</td>
<td>FG domain plus 1–800</td>
<td>49</td>
<td>52/34</td>
<td>18</td>
<td>Natively unfolded</td>
<td>Kap binding (1–640)</td>
<td>5, 9</td>
</tr>
<tr>
<td>Cluster I 801–952 84 41/46 −5</td>
<td>β-Sheet (814–959)</td>
<td>5</td>
<td>NPC anchor</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nup116</td>
<td>FG domain plus 1–960</td>
<td>61</td>
<td>57/30</td>
<td>27</td>
<td>Natively unfolded</td>
<td>Kap binding (165–715); Gle2 binding (110–166); NPC anchor (706–1113)</td>
<td>5, 9, 23</td>
</tr>
<tr>
<td>Cluster I 962–1111 82 46/42 4</td>
<td>β-Sheet (967–1113)</td>
<td>4</td>
<td>Natively unfolded</td>
<td>26, 28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nNup145a</td>
<td>FG domain plus 1–433</td>
<td>67</td>
<td>53/32</td>
<td>21</td>
<td>Natively unfolded</td>
<td>Kap binding (1–603)</td>
<td>9</td>
</tr>
<tr>
<td>Cluster I 434–605 81</td>
<td>β-Sheet (458–605)</td>
<td>10</td>
<td>NPC anchor (247–606); autoproteolysis (398–613)</td>
<td>28, 54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nsp1</td>
<td>FG domain plus 1–617</td>
<td>51</td>
<td>63/22</td>
<td>41</td>
<td>Natively unfolded (1–603)</td>
<td>Kap binding (1–603)</td>
<td>9, 55</td>
</tr>
<tr>
<td>Cluster I 618–809 91 43/40 3</td>
<td>Coiled coils (630–823)</td>
<td>3</td>
<td>NPC anchor (591–823)</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nup53</td>
<td>Cluster I 8–177 65 56/32 24</td>
<td>Natively unfolded</td>
<td>Molecular interaction Flexible linker or spacer</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster II 254–357 72 48/42 6</td>
<td>Natively unfolded</td>
<td>Flexible linker or spacer</td>
<td>Kap121 binding (401–448); membrane binding (449–475)</td>
<td>24, 56</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Clusters II and III 358–400 24</td>
<td>60/28 32</td>
<td>Natively unfolded</td>
<td>Amphilic α-helix (459–475)</td>
<td>24, 56</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster III 401–475 88 55/39 16</td>
<td>Amphilic α-helix</td>
<td>16</td>
<td>Molecular interaction Homodimerization</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nup59</td>
<td>N terminus 1–80 36 56/32 24</td>
<td>Natively unfolded</td>
<td>RRM fold (265–394)</td>
<td>24, 56</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster I 81–428 66 44/39 5</td>
<td>RRM fold (265–394)</td>
<td>5</td>
<td>Molecular interaction Homodimerization</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
regression: $dN = -0.0014(AACB) + 0.076$, $R^2 = 0.026$). In addition, the mean $dN$ value for the set of folded proteins is $0.06 \pm 0.01$, more than 5 times lower than the mean substitution rate for the FG Nups (Fig. 1A). Finally when the entire S. cerevisiae proteome was binned into quartiles based on AACB value, the 25% of proteins most biased for disorder-associated AAs show a significantly higher substitution rate than proteins near the median AACB value (data not shown). This correlation between disorder-associated AA composition and high substitution rate held true even when the data were controlled for the effects of translation rate, codon usage, and protein dispensability (data not shown). Thus, high substitution rates appear to be a general feature of disordered proteins, and this may explain the unusually rapid evolution rates of the FG Nups.

**DISCUSSION**

**Structural Disorder in FG Nups**—Our analyses of the AA compositions of Nups from the yeast, worm, fly, and human proteomes suggest that structural disorder is a conserved feature of the FG Nups in distantly related eukaryotes. Their unusually high content of charged and polar AAs (Table I and Supplemental Table S3) is a hallmark of disordered proteins in general (11). Biophysical experiments will be necessary to test conclusively the structural characteristics of each FG Nup in higher eukaryotes, but recent studies support our prediction that their FG domains also exist in a natively unfolded state (43–45). The apparent conservation of structural disorder in FG domains of Nups throughout Eukaryotae indicates that their inherent structural flexibility plays an important role in NPC function possibly for efficient capture and release of karyopherin-cargo complexes during transport across the NPC or as architectural elements of a size-selective permeability barrier.

The Natively Unfolded FG Domains of Nups Have Evolved Rapidly—We observed that the AA sequences of Saccharomyces FG Nups exhibit rapid evolution rates compared with the majority of yeast proteins, including non-FG Nups, Kaps, and a set of structured proteins (Figs. 1A, 3, and 5 and Supplemental Fig. S6). The FG domains are highly substituted (Fig. 1B) yet are apparently constrained to retain at least two features: (i) FG motifs (Fig. 2) and (ii) a high density of polar and/or charged AAs surrounding the FG motifs (Table I and Supplemental Table S3). These observations define a basic FG domain prototype that consists of FG motifs separated by 10–20 hydrophilic AAs. Indeed it appears that the intervening sequences between FG motifs are permissive to AA substitutions as long as the physicochemical properties of the re-

**TABLE II—continued**

<table>
<thead>
<tr>
<th>Nup domains defined by AA conservation</th>
<th>Domain AAs</th>
<th>Percent AA conservation</th>
<th>AA composition (percent disorder/order)</th>
<th>AACB value</th>
<th>Actual or predicted structure (AAs)</th>
<th>Actual or predicted function (AAs)</th>
<th>Refs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Clusters I and II</td>
<td>429–494</td>
<td>22</td>
<td>42/41</td>
<td>1</td>
<td>Structured</td>
<td>Structural bridge or spacer</td>
<td></td>
</tr>
<tr>
<td>Cluster II</td>
<td>495–528</td>
<td>79</td>
<td>50/38</td>
<td>12</td>
<td>Structured</td>
<td>Membrane binding</td>
<td></td>
</tr>
<tr>
<td>Nup60 N terminus</td>
<td>1–117</td>
<td>39</td>
<td>56/31</td>
<td>25</td>
<td>Natively unfolded</td>
<td>Kap123 binding</td>
<td>57</td>
</tr>
<tr>
<td>Cluster I</td>
<td>118–301</td>
<td>77</td>
<td>58/31</td>
<td>27</td>
<td>Natively unfolded</td>
<td>Molecular interaction</td>
<td></td>
</tr>
<tr>
<td>Between Cluster I and Box 12</td>
<td>302–349</td>
<td>33</td>
<td>64/25</td>
<td>39</td>
<td>Natively unfolded</td>
<td>Flexible linker or spacer</td>
<td></td>
</tr>
<tr>
<td>Box 12</td>
<td>350–386</td>
<td>89</td>
<td>70/22</td>
<td>48</td>
<td>Natively unfolded</td>
<td>Molecular interaction</td>
<td></td>
</tr>
<tr>
<td>FG domain plus</td>
<td>387–539</td>
<td>39</td>
<td>56/32</td>
<td>24</td>
<td>Natively unfolded</td>
<td>Kap95/60 binding</td>
<td>9, 57</td>
</tr>
<tr>
<td>Nup2 Cluster I</td>
<td>1–51</td>
<td>88</td>
<td>57/20</td>
<td>37</td>
<td>Natively unfolded</td>
<td>Kap60 binding (26–51)</td>
<td>31, 32</td>
</tr>
<tr>
<td>FG domain plus</td>
<td>52–600</td>
<td>41</td>
<td>59/25</td>
<td>34</td>
<td>Natively unfolded</td>
<td>Kap binding (186–561)</td>
<td>8, 58</td>
</tr>
<tr>
<td>Cluster II</td>
<td>601–708</td>
<td>82</td>
<td>51/34</td>
<td>17</td>
<td>$\beta$-Sheet ([383–720])</td>
<td>Gsp1/Ran binding</td>
<td>30, 57</td>
</tr>
<tr>
<td>Nup1 Cluster I</td>
<td>1–85</td>
<td>46</td>
<td>53/31</td>
<td>22</td>
<td>Structured</td>
<td>Molecular interaction</td>
<td></td>
</tr>
<tr>
<td>FG domain plus</td>
<td>121–1076</td>
<td>39</td>
<td>51/26</td>
<td>31</td>
<td>Natively unfolded</td>
<td>NPC anchor (4–212)</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>86–120</td>
<td>86</td>
<td>52/43</td>
<td>9</td>
<td>Natively unfolded</td>
<td>Kap binding (432–816);</td>
<td>9, 21, 58, 60</td>
</tr>
<tr>
<td></td>
<td>121–1076</td>
<td>39</td>
<td>51/26</td>
<td>31</td>
<td>Natively unfolded</td>
<td>Kap60 binding (1059–1076); Kap95 binding (975–987 and 1004–1009)</td>
<td></td>
</tr>
</tbody>
</table>
region are hydrophilic. Thus, the disordered structures of the FG domains may not require a specific sequence between FG motifs but rather a simple preponderance of hydrophilic AAs that promote extensive solvent interactions and inhibit the formation of stable secondary structure. Only AA sequences that contact interacting proteins, such as the FG motifs that bind to Kaps, are under selective pressure to be conserved (Table II).

To explain the high AA substitution rates of FG Nups, we examined several parameters known to influence protein evolution (Fig. 4). We concluded that neither the protein expression level, nor the interactivity, nor the essentiality of the FG Nups adequately explains their rapid evolution. Instead, a “structural disorder content” parameter showed a significant correlation with protein evolution rate, suggesting that structural disorder permits the high rate of AA substitution in FG Nups and possibly other natively unfolded proteins. High protein evolution rates have been observed for other disordered proteins (42). Importantly this suggests that the structural properties of a protein, in addition to its function, expression level, and contribution to organism viability, can influence its AA substitution rate.

Islands of AA Sequence Conservation in the Natively Unfolded FG Domains of Nups—Given their lack of stable structure, it is possible that the FG domains display their protein binding determinants as compact, linear sequences like the binding sites for proteins in DNA elements. Many small, discrete islands of AA sequences from 6 to 11 AA in length have been conserved in the FG domains of Nups during yeast evolution (Fig. 3 and Supplemental Fig. S6). These sequences often contain a single FG motif, although some conserved sequences intervene two FG repeats (Supplemental Fig. S7). These FG motifs may interact with different Kaps, of which there are at least 15 in yeast, or may specify different binding affinities for them.

Conserved Domains of Structure and Function in the FG Nups—We found that all previously characterized structural and functional domains in the *S. cerevisiae* FG Nups are highly conserved (Table II) despite the overall high AA substitution rates in these Nups. Notable conserved regions include the NPC anchoring domains of Nsp1, Nup42, and Nup116; the β-propeller structure that binds Dbp5 in the N terminus of Nup159; the Gde2 binding sequence (GLEBS) domain in Nup116; the Ran binding domain in Nup2; the high affinity Kap95 binding site in the C terminus of Nup1; the predicted RRM fold in Nup53 and Nup59; and the high affinity Kap60 binding sites in Nup2 and Nup1 (Table II). This suggests that other conserved, yet uncharacterized domains in FG Nups

**Fig. 4.** Percentile rank of each FG Nup within the *S. cerevisiae* proteome for various protein evolution parameters: substitution rate (dN), CAI, translation rate (TR) (proteins/s), and protein interactivity (INT) (binding partners/protein). Each symbol represents the percentile rank of one FG Nup. Data were not available for every FG Nup in some parameters.

**Fig. 5.** The relationship between AA composition and AA substitution rate in *Saccharomyces* Nups. The AACB value is the percent frequency of disorder-associated AAs (Ala, Arg, Gln, Glu, Gly, Lys, Pro, and Ser) subtracted by the percent frequency of order-associated AAs (Asn, Cys, Ile, Leu, Phe, Trp, Tyr, and Val) in a protein. Proteins with a high content of disorder- or order-associated AAs have positive or negative AACB values, respectively. A, the AA substitution rate (dN) of *Saccharomyces* Nups (FG and non-FG Nups) correlates with their AACB (linear regression: dN = 0.0065(AACB) + 0.168, R² = 0.369). Substitution rates were determined as in Fig. 1A. B, *Saccharomyces* proteins with highly significant homologs in the Protein Data Bank (PDB) of NMR and crystal structures were analyzed as in A (linear regression: dN = −0.0014(AACB) + 0.076, R² = 0.026). The dashed line indicates the median AACB value for the *S. cerevisiae* proteome.
may have important functions and structures. Based on that premise, our analysis predicts specific AA boundaries for novel structural and/or functional domains in FG Nups and offers refined AA boundaries for known NPC anchoring domains and Kap or Nup binding sites (Table II).

Using AA Sequence Conservation and AA Composition Data to Predict Domain Structure and Function in Natively Unfolded Proteins—For the yeast FG Nups, there appear to be four different combinations of AA composition and AA substitution rates that produce different structure-function predictions for protein domains. First, a high content of order-associated AAs (AACB value ≤19) and a low AA substitution rate (≥65% AA sequence conservation) predict a folded domain with a conserved molecular interaction; all domains listed as clusters in Table II fall under this category except for Cluster I in Nup60 and Nup53, which have higher AACB values. Second, a high content of order-associated AAs and a high AA substitution rate (≤35% AA conservation) predict a folded domain that may function mainly as a structural bridge or spacer between two domains; the regions between Clusters I and II in Nup59 and between Clusters II and III in Nup159 exemplify this. Third, a high content of disorder-associated amino acids (AACB value >19) and a low AA substitution rate predict a domain that is natively unfolded and participates in molecular interactions; Cluster I in Nup60 and Nup2 and Box12 in Nup60 are clear examples. Fourth, a high content of disorder-associated AAs and a high AA substitution rate predict a natively unfolded domain with molecular interaction sites that are limited to discrete islands of conserved AA sequences; all of the FG domains of Nups and the N termini of Nup1, Nup60, and Nup59 fall under this category. A similar domain, but without discrete islands of conserved AAs, may function as an unstructured flexible linker or spacer between two domains; the AA sequences between clusters in Nup53 and between Cluster I and Box 12 in Nup60 fall under this category.

In summary, we compiled more than 40 experimentally characterized or computationally predicted structural and functional domains in the FG Nups (Table II), and in each case, the local AACB value and the AA conservation value accurately match the local presence or absence of protein structure or function. The overall success of this analysis gave us confidence to make novel predictions regarding the structure and function of uncharacterized domains; our predictions are listed above and are highlighted in bold letters in Table II. Detailed structural, biochemical, and genetic characterizations of these FG Nup domains will be needed to further validate the predictive power of this approach.

Acknowledgments—We are grateful to Dr. Aaron Hirsh for providing the Saccharomyces substitution rate dataset prior to its publication, and we thank Drs. Dmitri Petrov and Andrey Sidow for helpful conversations. We also thank Samir Patel, Jessica Wright, and anonymous reviewers for comments on the manuscript.

* This work was supported in part by National Institutes of Health Grant GM061900 (to M. R.) and a National Science Foundation predoctoral fellowship (to D. P. D.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

REFERENCES


The Rapid Evolution of FG Nucleoporins

Functional profiling of the Saccharomyces cerevisiae genome. Nature 418, 387–391


41. Zhang, H., Ymer 255p. EMBO J. 18, 5781–5777


