Intrinsic Protein Disorder, Amino Acid Composition, and Histone Terminal Domains*

Published, JBC Papers in Press, November 21, 2005, DOI 10.1074/jbc.R500002200

Jeffrey C. Hansen, Xu Lu, Eric D. Ross, and Robert W. Woody
From the Department of Biochemistry and Molecular Biology, Colorado State University, Fort Collins, Colorado 80523

Core and linker histones are the most abundant protein components of chromatin. Even though they lack intrinsic structure, the N-terminal “tail” domains (NTDs) of the core histones and the C-terminal tail domain (CTD) of linker histones bind to many different macromolecular partners while functioning in chromatin. Here we discuss the underlying physicochemical basis for how the histone terminal domains can be disordered and yet specifically recognize and interact with different macromolecules. The relationship between intrinsic disorder and amino acid composition is emphasized. We also discuss the potential structural consequences of acetylation and methylation of lysine residues embedded in intrinsically disordered histone tail domains.

The core (H2A, H2B, H3, H4) and linker (H1 family) histones make up the fundamental protein components of chromatin fibers (1, 2). The N-terminal “tail” domains (NTDs) of the core histones and the C-terminal tail domain (CTD) of linker histones are intrinsically disordered, yet they are able to bind to many different macromolecular partners in chromatin. For example, the histone H3 and H4 NTDs interact with sites on other nucleosomes during chromatin condensation (3, 4) and bind to proteins such as Sir3p (5) and p900 (6). The H1 CTD interacts with linker DNA in a chromatin fiber (1, 2) and also binds to proteins such as DFF/40CAD (7). This article focuses on the roles of intrinsic protein disorder in histone function. We highlight recent findings indicating that amino acid composition is the key determinant of molecular recognition by the histone tail domains and other intrinsically disordered protein regions. We also discuss how acetylation and methylation of lysine residues may modulate macromolecular interactions by altering the local physicochemical properties of intrinsically disordered histone domains.

What Is Intrinsic Disorder?

Proteins (or sizeable regions of proteins) that lack a well defined conformation under native conditions are referred to as “intrinsically disordered.” Many intrinsically disordered proteins (IDPs) are functional, adopting a well defined conformation upon interacting with a target molecule. Thus, the principle that protein function requires a well defined conformation must be modified; an isolated protein need not have a unique conformation, but the protein-target complex must. A corollary is that if the protein in question interacts with more than one target, it may adopt a corresponding number of different conformations. One of the more surprising aspects of IDPs is their ubiquity, especially in eukaryotes. The most rigorous analysis to date predicts that long IDP regions are found in an average of 33% of eukaryotic proteins but in only 2% of archaeal and 4% of eubacterial proteins (8). Conformational adaptability is generally considered to be one of the driving forces for the evolution of IDPs.

IDPs have several features that distinguish them from classical globular proteins. Experimentally, IDPs are recognized by a far-UV CD spectrum characteristic of unordered proteins: sharp peaks in NMR, low dispersion of chemical shifts, negative 1H-2H nuclear Overhauser effects, a radius of gyration or hydrodynamic radius comparable with that of the protein in concentrated urea or guanidinium chloride, and a marked susceptibility to proteases (9, 10). The absence of order in a crystal structure often is indicative of an intrinsically disordered domain as well.

IDPs can be predicted from amino acid sequence data with good accuracy. In one study of more than 900 nonhomologous proteins, predictions of disordered regions more than 40 residues in length gave less than 6% false positives (11). Predictive algorithms score sequences according to the flexibility, hydrophathy, charge, and other physicochemical properties of the amino acid residues (12–15).

Compositional bias is a common feature of IDPs (9, 14). Some amino acids are substantially more abundant in IDPs than in the average folded protein, whereas others are rare or absent. The bias generally favors hydrophilic amino acids and discriminates against hydrophobic residues. Thus, IDPs are generally rich in Arg, Gln, Glu, Lys, Pro, and Ser. They are deficient in Cys, Ile, Leu, Phe, Trp, Tyr, and Val. The other amino acids are present at levels comparable with those in the average folded protein (Met, Thr) or are enriched in some IDPs and depleted in others (Ala, Asn, Asp, Gly, His). As will be discussed below, there is now compelling evidence indicating that the relationship between amino acid composition and IDPs is far more complex than the simple trends described above.

The compositional bias of IDPs accounts for their inability to fold; the paucity of hydrophobic groups precludes the formation of a hydrophobic core about which the chain can fold. Further, many IDPs have a large excess of basic or acidic amino acids and hence are highly charged at neutral pH. The charge on such proteins acts to destabilize a compact structure. Interaction with target proteins or nucleic acids overcomes these problems, allowing the IDP to undergo a concerted folding-binding process. Hydrophobic groups of the IDP are buried in the IDP-target interface, interacting with exposed hydrophobic groups of the target. The target usually has a charge opposite to that of the IDP, at least locally, leading to a lower charge density in the complex.

What advantages of IDPs account for their abundance, especially in eukaryotes?

1) The coupling of folding and binding provides enhanced specificity at the expense of binding affinity (16, 17). The negative ΔS of folding is paid for by a large negative ΔH of binding. 2) The binding energy for an IDP is more favorable than for a compact protein with the same number of residues because the area of the interface is substantially larger for the IDP (18). 3) The flexibility of an IDP allows it to bind a number of different targets. 4) Flexibility also permits more rapid binding to the target through the mechanism of “flycasting” (19). An IDP is extended and thus presents multiple points of attachment for the biunimolecular step of encounter complex formation. The encounter complex can then undergo rapid unimolecular steps to the stable complex. 5) Flexibility permits ready access of a side chain to modify enzymes and to targets that recognize the modification. 6) A flexible, extended structure can be rapidly degraded by intracellular proteases, providing a facile pathway for down-regulation. Susceptibility to proteases is also a potential disadvantage for IDPs. Survival in the cell implies that IDPs must be complexed to targets most of the time.

The coupled process by which an IDP folds and binds to its target bears some resemblance to the induced fit concept of enzyme-substrate binding and allostery. However, in induced fit, ligand binding perturbs an equilibrium between two compact, well defined protein conformations, whereas binding of an IDP to a target involves a disorder → order transition of the IDP concomitant with formation of a macromolecular complex. The target may have a compact, well defined conformation, or it may be an IDP itself. In the final state, the IDP-target complex has a compact structure with classical protein folds (at least as a core), possibly with one or more flexible appendages.

Most of the functions of IDPs are related to molecular recognition of DNA, RNA, and other proteins. Fully or partially disordered proteins are especially common in processes such as transcription, cell cycle regulation, signal transduction, and chaperoning of the folding of proteins and RNA (20–22). Partially disordered regions are commonly found at the amino and carboxyl ends of proteins but can be present at internal sites as well. IDPs have been grouped into two main categories based on function: mediators of macromolecular interactions and entropic connectors/springs (20). Because of their ubiquity, other functions are likely to be identified as well.

Intrinsic Disorder, Amino Acid Composition, and Linker Histone CTD

Linker histones comprise a family of nucleosome-binding proteins that stabilize condensed chromatin and regulate genome function (1, 2, 23). The linker
histones of most eukaryotes have a very simple domain organization, consisting of a central winged helix fold, a short N-terminal extension, and a long basic C-terminal domain (Fig. 1). Little is known about the NTD region. The winged helix domain interacts with nucleosomes (1, 2). The CTD is ~100 residues in length, enriched in Lys, Ala, and Pro, and unstructured in aqueous solution (24). The determinants required to stabilize chromatin fibers in highly condensed conformations lie in the CTD (25, 26).

There are six somatic linker histone isoforms in most higher eukaryotes. Although the primary sequence of the isoform CTDs has diverged (24), the amino acid composition of the CTDs is surprisingly similar (Table 1). Each of the CTDs consists of ~40% Lys, ~20%–35% Ala, and ~15% Pro, Ser, Thr, Gly, and Val are present in all isoform CTDs in smaller, variable amounts. His, Tyr, Trp, Met, and Cys are never found in any of the isoform CTDs, and the other seven amino acids are sporadically present once or twice in a particular CTD. Val is the only hydrophobic amino acid found in all CTDs. The characteristic amino acid composition of the linker histone CTDs suggests that this domain functions as an IDP region. Recent experimental evidence supports this idea and has focused attention on the relationship between intrinsic disorder and amino acid composition. CTD truncation mutants were used to define the location of the amino acid residues involved in mouse H1 CTD function during chromatin condensation (26). The determinants for both linker DNA binding and chromatin fiber stabilization were localized to two distinct, separated regions of the CTD (Fig. 1). The functional regions are somewhat enriched in Val, but otherwise the amino acid composition of all CTD regions examined was similar. The two functional CTD regions can be interchanged, even though their primary sequences are different. This suggests that the key properties involved in DNA binding and chromatin condensation are amino acid composition and location of the CTD region relative to the winged helix domain, not primary sequence.

The H1 CTD also has been shown to mediate the protein-protein interactions involved in H1-dependent activation of the apoptotic nucleosome, DFF40/CAD (7). The CTD region that binds to the enzyme is large and partially overlaps with the two CTD regions that bind linker DNA and stabilize condensed chromatin (Fig. 1). Interestingly, all somatic linker histone isoforms activated the enzyme identically in vitro. Moreover, all free CTD peptides that were at least 47 residues in length could bind to and activate the enzyme, regardless of their primary sequence and original location in the intact CTD. Thus, amino acid composition and location of the CTD region relative to the winged helix domain also appear to be the determinants of CTD-protein interactions. Together, the studies of linker histone CTD involvement in chromatin condensation and DFF40/CAD activation demonstrate that the H1 CTD is an IDP region capable of interacting with both DNA and proteins and suggest that CTD function is linked to a distinctive amino acid composition.

Intrinsic Disorder, Amino Acid Composition, and Core Histone NTDs

The functions of the core histone NTDs have been investigated extensively (2–4, 27). These domains currently are of particular interest because specific patterns of NTD acetylation and methylation regulate gene expression and other nuclear processes (28–30). The NTDs are not observed in the crystal structures of the nucleosome (31). Free NTD peptides are disordered (see Ref. 27). In nucleosomes, the NTDs adopt increased α-helical content when bound to DNA (32, 33). All four of the core histone NTDs participate in the internucleosomal interactions that drive chromatin fiber condensation (3, 4). In addition, the H3 and H4 NTDs also bind to proteins such as Sir3p and p300 (5, 6). The amino acid composition of the core histone NTDs is shown in Table 1. The NTDs have a low percentage of hydrophobic residues and are highly enriched in Lys, Gly, and Arg residues. By all available criteria, the core histone NTDs also possess the characteristics of an IDP region. Unlike the linker histone CTDs, their primary sequences are highly conserved.

**Does Amino Acid Composition Define Different Types of Functional IDP Regions?**

A closer examination of the amino acid composition of the core histone NTDs reveals several interesting trends (Table 1). The composition of the H2A and H4 NTDs is very similar but differs significantly from that of the linker histone CTDs. Specifically, the H2A and H4 NTDs have no Pro, more Gly and Arg, and fewer Ala than the linker histone CTDs. On the other hand, the amino acid composition of the H2B and H3 NTDs is surprisingly similar to that of the linker histone CTDs (Table 1). Based on amino acid composition, at least two different types of IDP regions are involved in histone function.

It is of note that the characteristic amino acid composition of the H1 CTDs also is found in other proteins. A region of 38 residues in the core histone variant, macroH2A, has a very similar amino acid composition as the linker histone CTDs (Table 1). However, in this case, the IDP region is located internally and connects two structured domains (Fig. 1). The amino acid composition of the macroH2A connector domain suggests that it is an IDP region. The broader implication is that the linker histone CTD actually is a specific type of IDP region that is found in different locations within different proteins.

Further support for the existence of specific types of IDP regions comes from studies of yeast prions (infectious proteins). The yeast prion proteins Ure2p and Sup35p each contain a N-terminal prion domain that is sufficient for prion formation but dispensable for the normal function of the protein (34). In both cases the prion domains are intrinsically disordered, but upon conversion to the prion conformation, they self-associate to form self-propagating amyloid-like fibrils (35). The prion conformation is a folded β-domain structure, as with other amyloid-forming proteins. Randomizing the primary sequence of the Sup35p and Ure2p prion domains while keeping the amino acid composition constant does not inhibit prion formation (36, 37), indicating that amino acid composition is the fundamental determinant of amyloid formation in these systems and not the primary sequence. The amino acid composition of the prion domains is consistent with that of an IDP but differs significantly from that of the H1 CTDs (Table 1). In particular, Ure2p and Sup35p are highly enriched in Asn and Gln rather than Lys, Ala, and Pro. An intriguing possibility is that amino acid composition determines which type of secondary structure is formed when an IDP region binds to a target, e.g. Asn/Gln-rich IDP regions may form β-domains, whereas Lys/Ala/Pro IDP regions may form α-helices. (Although Pro is generally considered to be a strong helix
In summary, the relationship between amino acid composition and IDP function is complex. The same is true for the relationship between amino acid composition and primary sequence. Using amino acid composition as a criterion, it appears that there are many different types of functional IDP regions just as there are many different types of functional protein folds.

**Lysine Acetylation and Methylation in IDP Regions**

Lysine residues within the intrinsically disordered core histone NTDs are modified through addition of methyl or acetyl groups. Specific patterns of NTD acetylation and methylation are involved in the regulation of transcription, replication, and other nuclear processes (28–30). These patterns of modifications often function by establishing or disrupting specific binding surfaces for other proteins. For example, the proteins HP1 (39, 40) and polycomb (41) can only bind to an H3 NTD peptide if the peptide is di- or trimethylated on Lys-9. A question that has remained largely unanswered at the molecular level is how acetylation and methylation influence the ability of the core histone NTDs to participate in specific protein-protein interactions.

Acetylation and methylation both affect the charge density, size, and hydrophobicity of the Lys side chain. Hydrophobicity may be particularly important because there are very few hydrophobic amino acids in IDP regions. Acetylation of Lys makes formation of secondary structures more favorable by decreasing the positive charge density and enhancing hydrophobic character. The free charged NH group is converted into a neutral amide linkage capped with a hydrophobic methyl group. Methylation of Lys leaves the positive charge density unaltered but replaces up to three polar NH groups capable of hydrogen bonding with hydrophobic methyl groups. Acetylation and methylation of Lys ultimately create unique amino acids with unusual properties. Hence we do not believe that acetylation and methylation simply create patterns of “marks” that are recognized by other proteins. Rather, we feel that acetylation and methylation alter the fundamental IDP properties of the NTDs as a prerequisite for coupled NTD folding and target binding.

This view is supported by the finding that nonspecific hyperacetylation of the core histone NTDs increases their average a-helical content (33). Moreover, Dion et al. (42) have shown that H4 acetylation at Lys-5, -8, and -12 functions in vivo through a nonspecific, cumulative mechanism. X-ray and NMR studies of complexes between methylated H3 peptides and the HP-1 and polycomb chromodomains have provided insight at the molecular level (39–41). In all such complexes, the disordered methylated H3 peptide adopts an extended chain structure, actually serving to fill in a /H9252-sheet in several cases. The extended chain structure optimizes interactions of both the backbone and side chain groups of the peptide with those of the protein. Importantly, modified Lys residues are recognized by specific features. For example, chromodomains have three aromatic side chains that form a hydrophobic cage that interacts with the methyl group(s) of the methylated Lys side chain. Without the hydrophobic tripod by the methyl groups, binding would not be possible and the NTD peptide would not assume the extended chain secondary structure.

### Concluding Remarks

The biological need for the core histone NTDs and linker histone CTD to interact with many modifying enzymes and recognition modules with widely varying structures can be readily accommodated if these domains are intrinsically disordered. We envision that the histone terminal domains interact with their targets through several different modes. In many cases, they bind as extended chains to sites that recognize the local sequence properties as in the case of the recognition motifs discussed in the previous section. In other cases, these IDP regions can fold into a-helical segments, /-hairpins, or other simple motifs, burying hydrophobic groups introduced by modifications and/or in combination with hydrophobic groups on the target. If binding depends primarily on the stability of the secondary structure formed, it may be more important to conserve amino acid composition rather than primary sequence. In this regard, the sequence conservation of the core histone NTDs may be related to maintaining unique post-translational modification sites more so than the primary sequence per se.

The IDP regions of linker histones and yeast amyloid proteins challenge the paradigm that the primary amino acid sequence and corresponding main and side chain interactions dictate formation of a unique local polypeptide conformation with the lowest free energy state. In these systems, a specific amino acid composition is conserved and correlated with protein function, whereas the primary sequence varies. A recent study of 718 IDP sequences using support vector machine analysis (43) concluded that amino acid composition is the only parameter needed to accurately recognize IDPs and that IDP regions are defined by physical properties of a short stretch of amino acids rather than the interactions dictated by the primary sequence of amino acids. Evidence is mounting that in many cases there is a direct correlation between amino acid composition, intrinsic disorder, and protein function. Even in situations where the primary sequence is conserved (such as the core histone NTDs), local amino acid composition may be the key property required for molecular recognition. Examination of the relationships between amino acid composition and IDP function in a wide range of biological systems is likely to reveal new principles of protein structure and molecular recognition.

### REFERENCES

1. Wolffe, A. (1998) Chromatin: Structure and Function, 3rd Ed., Academic Press, San Diego
2. Hansen, J. C. (2002) Annu. Rev. Biophys. Biomol. Struct. 31, 361–392
3. Dorigo, B., Schulch, T., Bystrick, K., and Richmond, T. J. (2003) J. Mol. Biol. 327, 85–96
4. Gordon, F., Luger, K., and Hansen, J. C. (2005) J. Mol. Biol. 328, 23071–23076
5. Hecht, A., Laroche, T., Strahl-Bolsinger, S., Gasser, S. M., and Grunstein, M. (1995) Cell 80, 583–592
6. An, W., and Roeder, R. G. (2003) J. Mol. Biol. 278, 1504–1510
7. Wdowak, P., Kalinowska, M., Parsengeh, M. H., Lu, X., Hansen, J. C., and Garrard, W. T. (2005) Biochemistry 44, 7871–7878
8. Ward, J. J., Sodhi, S. J., Mcguflin, L. J., Buxton, B. F., and Jones, D. T. (2004) J. Mol. Biol. 337, 635–645
9. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipsps, K. W., Ausiio, J., Nissen, M. S., Reeves, R., Kang, C., Kissingar, C. R., Bailey, R. W., Grisswold, M. D., Chiu, W., Garner,
MINIREVIEW: Intrinsic Disorder in Histone Function