INTRODUCTION

Paramyxovirinae, which include major human pathogens such as parainfluenza virus and measles virus (MV), are enveloped viruses with a non-segmented, negative, single-stranded RNA genome encapsidated by the nucleoprotein (N) within a helical nucleocapsid. Transcription and replication are carried out on this (N : RNA) template by a viral RNA-dependent RNA polymerase complex, made of the phosphoprotein (P) and the large protein (L) (reviewed by Lamb & Kolakofsky, 2001). Association of P with the phosphoprotein (P) and the large protein (L) (reviewed by Lamb & Kolakofsky, 2001). Association of P with the soluble, monomeric form of N (N\(^\circ\)) prevents its illegitimate self-assembly onto cellular RNA. The assembled form of N (N\(^{\text{NUC}}\)) also forms complexes with P and P–L during transcription and replication (Lamb & Kolakofsky, 2001).

N consists of two regions: an N-terminal moiety, well conserved in sequence, N\(^{\text{CORE}}\), and a hypervariable, C-terminal moiety, N\(^{\text{TAIL}}\). N\(^{\text{CORE}}\) contains all the regions necessary for self-assembly and RNA binding. N\(^{\text{TAIL}}\) binds P within both N\(^{\text{NUC}}\) and N\(^\circ\) and is required for N:RNA to act as a template for viral RNA synthesis (Bankamp et al., 1996; Buchholz et al., 1994; Curran et al., 1993; Harty & Palese, 1995; Nishio et al., 1999).

From a structural point of view, P is the best-characterized protein of the replicative complex. P is organized into two moieties that are functionally and structurally distinct: a C-terminal moiety (PCT) and an N-terminal moiety (PNT).

PCT is the most conserved in sequence and contains all regions required for virus transcription, whereas PNT, which is poorly conserved, provides several additional functions required for replication (Curran & Kolakofsky, 1999). P forms oligomers through a coiled-coil motif located within PCT. PCT also contains the region responsible for binding to L (Liston et al., 1995; Smallwood et al., 1994), as well as the regions necessary for binding N\(^{\text{NUC}}\) (Harty & Palese, 1995; Ryan et al., 1991). The extreme C-terminal domain of PCT (called XD, 'X domain') is responsible for binding to N\(^{\text{NUC}}\), as well as for stable binding to N\(^\circ\) (Curran et al., 1995b; Nishio et al., 1996; Shaji & Shaila, 1999). The structure of the Sendai virus (SeV) P multimerization domain (PMD) has been solved by X-ray crystallography. It is composed of a short bundle of \(x\)-helices located upstream of the coiled-coil (Tarbouriech et al., 2000b). The only structural information available on PNT concerns MV PNT, which is unstructured in vitro (Karlin et al., 2002b). PNT prevents the illegitimate self-assembly of N\(^\circ\) by binding to it. The main N\(^\circ\)-binding site has been mapped to the N terminus of PNT in all Paramyxovirinae (Curran et al., 1995b; Nishio et al., 1996; Precious et al., 1995; Shaji & Shaila, 1999; Tober et al., 1998). Beyond P, the P mRNA encodes a variety of proteins, including proteins consisting of either PNT alone (proteins W and R) or PNT fused to a zinc-binding region (protein V) (Lamb & Kolakofsky, 2001).

L is thought to be a multifunctional enzyme carrying most catalytic functions necessary for synthesis of viral RNA, such as RNA-dependent RNA polymerase (Poch et al., 1990;
proteins or protein domains that are unstructured intrinsically disordered (or natively unfolded) designate absence of a binding partner (reviewed by Dunker under physiological conditions of salt and pH, in the 2001; Uversky, 2002b; Wright & Dyson, 1999). In recent years, it has been discovered that they are usually distinguished from globular proteins by common sequence features. Intrinsically disordered proteins (IDPs) tend to have a low sequence complexity (i.e. they make use of fewer types of amino acids) (Romero et al., 2001). They are generally enriched in amino acids preferred at the surface of globular proteins (A, R, G, Q, S, P, E and K) (termed ‘disorder-promoting amino acids’) and are depleted in W, C, F, I, Y, V, L and N (‘order-promoting amino acids’) (Williams et al., 2001).

Their distinct sequence properties allow disordered regions to be predicted with good accuracy. A neural network-based predictor of naturally disordered regions (PONDR) allows predictions of long disordered regions (>40 aa) (LDRs) of proteins with very good confidence (>99.6%) (Li et al., 1999; Romero et al., 1997). Long ordered regions (>40 aa) are predicted with a similar confidence. PONDR, however, tends to underpredict disordered regions and therefore its disorder predictions can be considered as conservative (Dunker et al., 2002b).

Another quantitative method to characterize disordered proteins or protein domains relies on their mean net charge/mean hydrophobicity ratio, which is distinctly higher than that of their structured counterparts. This allows the two classes of proteins to be discriminated with a very good accuracy (Uversky, 2002b; Uversky et al., 2000). However, contrary to PONDR, the hydrophobicity/net charge method can only be applied to modular regions and therefore requires prior knowledge of the organization of the protein under study.

A more qualitative method is hydrophobic cluster analysis (HCA) (Callebaut et al., 1997). A HCA plot consists of a two-dimensional, helical representation of a protein sequence, allowing an intuitive visualization of clusters of hydrophobic amino acids (generally corresponding to secondary structure elements in globular proteins). Because the hydrophobic cluster information is plotted directly on the primary sequence, globular regions can be visualized, owing to their typical, thick distribution of hydrophobic clusters. On the contrary, non-globular regions are generally poor in hydrophobic residues and rich in polar residues.

Combining these different computational methods, we have analysed the presence and extent of structural disorder within Paramyxovirinae N, P and L. We focused on the three best-characterized genera (Morbillivirus, Respirovirus and Rubulavirus), calling in other viruses when they present informative differences. We used PONDR as a first guide to delineate disordered regions within N, P and L. Then, the precise boundaries of disordered regions were manually refined with HCA. This approach allowed the identification of modular regions to which we applied the net/charge hydrophobicity method. In our analysis, we have also taken into account other indicators of structural disorder, such as low sequence complexity (Romero et al., 2001), lack of predicted secondary structure (Liu et al., 2002) and sequence variability (Brown et al., 2002).

We show that spectacularly long unstructured regions are found in two (out of three) actors of the replicative complex of Paramyxovirinae, namely N and P. These disordered regions are conserved in the different genera, implying functional significance. Beyond providing a contribution to the study of the rather poorly investigated field of IDPs, the identification of disordered regions within these proteins facilitates their study at the structural and functional level.

**METHODS**

**Sequence retrieval.** Sequences for this study were obtained from the NCBI. Sequence accession numbers for P are: MV, CA919364; SeV, P04860; canine distemper virus (CDV), AA15481; Nipah virus (NiV), NP_112022; Menangle virus (MeV), AAK62280; Newcastle disease virus (NDV), NP_071467; Hendra virus (HeV), NP_047107; goose paramyxovirus (GPV), AAN04252; human parainfluenza virus type 2 (hPIV-2), NP_598402; hPIV-4, A43685; avian paramyxovirus (APV), NP_150058; simian virus type 5 (SV5), P11208; Tioman virus (TiV), NP_665865; mumps virus (MuV), P16072; and La-Piedad-Michoacan-Mexico virus (LFMV), AAL9693. Sequence accession numbers for N are: bovine parainfluenza virus type 3 (bPIV-3), AAF28254; MV, P35972; SeV, Q07097; and hPIV-2, P21737.

**Plotting mean net charge against mean hydrophobicity to assess whether a protein is intrinsically disordered.** The mean net charge (R) and the mean hydrophobicity (H) of a protein were calculated as described in Karlin et al. (2002b) and Uversky et al. (2000). For a given protein, R is then plotted against H. The charge/hydrophobicity diagram is divided into two regions by a line, which corresponds to the equation \( H = (R + 1.151) / 2.785 \). In the left part of the diagram [where \( H < (R + 1.151) / 2.785 \)], a protein is predicted as disordered, whereas it is predicted as ordered in the right part. The net charge/hydrophobicity method is only applicable to a protein (or protein region) provided it is not composed of shorter,
structurally independent modules. It might otherwise give conflicting results. It was only validated for regions >50 aa (Uversky et al., 2000). An estimation of its error rate can be drawn from Uversky (2002b). In that study, no globular protein was found to have a ratio located on the left side of the line, indicating that the positive error rate for the prediction of disordered proteins must be very low. However, five unfolded proteins out of 105 – which were all borderline – were wrongly assigned as being globular, indicating a negative error rate of about 5%.

**PONDR prediction of unstructured regions.** Sequences were submitted to the PONDR server (http://www.pondr.com/) using the default integrated predictor VL-XT (Li et al., 1999; Romero et al., 2001). The threshold for reliable (>99.6%) predictions of disorder, or of order, is set to 40 residues. Access to PONDR was provided by Molecular Kinetics (Pullman, WA, USA) under licence from the WSU Research Foundation. PONDR is copyright @1999 by the WSU Research Foundation, all rights reserved.

**HCA and amino acid composition analysis.** HCA was carried out with the program DRAWHCA (Callebaut et al., 1997). The average sequence composition of globular proteins was taken from Tompa (2002). If the average composition of an amino acid X in globular proteins is CGX, and CPX is the composition in X of a protein P, deviation from the composition in X of globular proteins was defined for P as (CPX—CGX)/CGX.

**Identification of low sequence complexity segments and secondary structure predictions.** Low sequence complexity segments were identified using the ‘low complexity filter’ of the BLAST program at the NCBI (http://ncbi.nlm.nih.gov), based on the program SEG (Wootton, 1994). Secondary structure predictions were performed with PSI-PRED (McGuffin et al., 2000) and the PREDICT protein server (Rost, 1996). The results presented are a consensus of both methods.

**Multiple sequence alignment of N and P.** The sequences of Paramyxovirinae Ncore were aligned using CLUSTALW and manually refined with SEAVIEW (Galtier et al., 1996). The sequences of NTAIL could not be aligned among different genera. The sequence alignment of Paramyxovirinae PCT was generated in the same way and was essentially the same as that reported in Curran et al. (1995a), with the exception of the central region of PCT, which is not presented in Curran et al. (1995a). The sequences of PNT among different genera could not be aligned, with the exception of the N-terminal regions of Rubulavirus, Henipavirus and Avulavirus PNT (see below).

**Multiple sequence alignment of the N-termini of PNT.** The sequences related to the N terminus of Rubulavirus PNT (aa 1–100) were retrieved by PSI-BLAST (Altschul et al., 1997) from SWISS-PROT (Bairoch & Apweiler, 2000), PDB (Berman et al., 2000) and translated GenBank (Benson et al., 2002). Fragments and duplicates were discarded. PSI-BLAST converged after seven iterations. The most distant hit is given with a significant (4 x 10^-5) E-value. All positive hits were used as subsequent PSI-BLAST queries and cross-validated.

As the sequences are not closely related, the first alignment was done using CLUSTALW (Thompson et al., 1994) with the slow algorithm, an identity matrix, a window of 3 aa and the standard gaps penalties. The alignment was manually refined with SEAVIEW (Galtier et al., 1996) using predicted structural information. The alignment was drawn using ESPript 2.0 (Gouet et al., 1999).

**RESULTS**

MV PNT and NTAIL are intrinsically disordered (Karlin et al., 2002b; Longhi et al., 2003). We have analysed the sequence properties of Paramyxovirinae N, P and L in order to determine whether such intrinsic disorder is a conserved feature in these viruses. The identification of disordered regions is expected to help decipher their modular organization and thereby facilitate their structural characterization.

**All Paramyxovirinae NTAIL are intrinsically disordered**

PONDR predicts at least one LDR (>40 aa) in NTAIL in the three genera but none in Ncore (Fig. 1A). The HCA plot of MV N shows the presence within Ncore of two large regions rich in hydrophobic clusters (Fig. 2). An antigenic region of Ncore (Giraudon et al., 1988) is clearly visible as a short interruption of hydrophobic clusters, indicating that it may form a loop exposed to the solvent. Conversely, MV NTAIL contains a strikingly long region totally devoid of hydrophobic clusters (aa 421–494) (Fig. 2), which correlates well with the PONDR prediction. In all Paramyxovirinae, NTAIL has little or no predicted secondary structure (shown for MV on Fig. 2). Furthermore, in the three genera, NTAIL possesses a combination of low hydrophobicity and relatively high net charge (due to the presence of numerous acidic residues) typical of IDPs (Fig. 3). Finally, NTAIL is also greatly variable in sequence. Thus, the sequence properties of Paramyxovirinae NTAIL converge to show that they are intrinsically disordered.

**A flexible linker between the coiled-coil and XD in all Paramyxovirinae?**

PONDR predictions point to the possible presence of a disordered linker between the coiled-coil and XD. In all respiroviruses, with the exception of SeV, this region is predicted to be a LDR (data not shown). In SeV, this region displays a pattern of predicted disorder alternating with borderline order (Fig. 1B, circled). A similar pattern of alternating order and disorder can be seen in all morbilliviruses (circled in Fig. 1B for CDV). In almost all rubulaviruses, the corresponding region is predicted as a LDR (shown in Fig. 1B for hPIV-2). HCA predictions are not conclusive but show that this region exhibits peculiar properties in all viruses. In particular, the linker region of SeV P, while not being as poor in hydrophobic clusters as LDRs found in P, contains distinctly fewer such clusters than globular regions, such as PMD or XD (Fig. 4). At the same time, it is enriched in disorder-promoting residues (data not shown). A region of analogous composition can be found between PMD and PX in Morbillivirus (data not shown) and Rubulavirus (shown in Fig. 5 for hPIV-2). Finally, contrary to Rubulavirus (Fig. 5) and Morbillivirus (data not shown), the linker of Respirovirus PCT is predicted to lack any secondary structure (shown in Fig. 4 for SeV).

In conclusion, a disordered linker is very probably found in Respirovirus and Rubulavirus PCT. It is probably found in Morbillivirus, too, but we could not reach the same degree of confidence as for our other predictions.
The central regions of other genera share little or no sequence similarity among them but all have the same peculiar composition (being rich in G, S and A). They are depleted in most ‘order-promoting’ residues and enriched in most ‘disorder-promoting’ residues (Fig. 6), suggesting that they might be disordered (Williams et al., 2001). In agreement with these sequence features, PONDR predicts a LDR in the central regions of Rubulavirus and Morbillivirus P (Fig. 1B). Furthermore, their HCA plots are typical of disordered regions, they lack predicted secondary structure (shown for hPIV-2 in Fig. 5) and contain low sequence complexity segments (Fig. 5). In both genera, their net charge/hydrophobicity ratios are those of globular proteins, although they are borderline (Fig. 3). In conclusion, the central region is likely intrinsically disordered. Interestingly, the central region of P overlaps the V ORF. Similarly, the C ORF overlaps PNT (Lamb & Kolakofsky, 2001), which is unstructured (see below). This suggests that the presence of unstructured regions might be a common feature of proteins encoded by overlapping reading frames.

**The PNT moiety**

There are functional and structural differences between Morbillivirus and Respirovirus PNT, which are acidic and 230–320 aa in length, and Rubulavirus PNT, which are shorter (about 160 aa), basic and have a short stretch of sequence similarity with the C proteins of Respiroviruses (Lamb & Kolakofsky, 2001). Moreover, Rubulavirus V is found in virions (Paterson et al., 1995) and binds RNA through a stretch of basic residues located within PNT (Lin et al., 1997), contrary to the V proteins of the other genera. Consequently, we have analysed separately Morbillivirus and Respirovirus PNT on one hand and Rubulavirus PNT on the other hand.

**Respirovirus and Morbillivirus PNT are disordered but are composed of two distinct regions**

We reported previously that Morbillivirus and Respirovirus PNT are predicted to be largely disordered using both PONDR (Fig. 1B) and the hydrophobicity/net charge method (Fig. 3) (Karlin et al., 2002b). Further analysis using HCA and secondary structure predictions reveals that PNT is in fact divided into two regions: an N-terminal region rich in hydrophobic clusters associated with a clear α-helical propensity followed by a region devoid of hydrophobic clusters and of predicted secondary structure (shown for SeV in Fig. 4). The α-helical propensity of the extreme N-terminal region of PNT is in agreement with data available in the literature on MV PNT (Karlin et al., 2002b).

**Modular organization of Rubulavirus PNT**

We found previously that Rubulavirus PNT has the hydrophobicity/net charge ratio typical of globular proteins (Fig. 3) (Karlin et al., 2002b). However, we present evidence indicating that Rubulavirus PNT is composed of at least two
Fig. 2. HCA plot of MV N. Conventions are given in the caption. Globular regions (framed) are characterized by a thick distribution of hydrophobic clusters, while unstructured regions are poor or devoid of hydrophobic clusters. LDR and predicted secondary structure elements are shown. An antigenic region of N_CORE (see text) (Giraudon et al., 1988) is highlighted above the diagram. There are no low complexity regions in MV N.
modular regions (see below). Since the hydrophobicity/net charge method can be applied only to modules, the predictions of globularity obtained on the whole PNT cannot be considered reliable. Using PSI-BLAST (Altschul et al., 1997), we have identified in Rubulavirus PNT a conserved N-terminal region with a previously unreported sequence identity with the N-termini of Henipavirus and NDV PNT (Fig. 7). Conversely, there is no detectable sequence identity among the corresponding regions of Morbillivirus and Respirovirus PNT. The conserved N-terminal region of Rubulavirus is distinguished by numerous hydrophobic clusters and by a high z-helix-forming potential (shown for hPIV-2 in Fig. 5). Notably, it has the hydrophobicity/net charge ratio typical of globular proteins (shown for hPIV-2 in Fig. 3). It contains an N\textsuperscript{\textprime}a-binding region and a nuclear localization signal, which can both function in isolation, arguing for some degree of functional independence of this region (Watanabe et al., 1996).

The region downstream of this conserved module is mostly disordered, as estimated by PONDR (Fig. 1B), HCA (Fig. 5) and as suggested by the lack of predicted secondary structure (Fig. 5). Nevertheless, two subgroups of Rubulavirus can be distinguished. The first subgroup includes hPIV-2 and closely related viruses such as SV5, SV41 and MuV, while the second comprises more distant viruses such as MeV, TiV or LPMV. In the first subgroup, the disordered region downstream of the conserved region contains a lysine/arginine-rich RNA-binding region (Fig. 5) (Lin et al., 1997), which is noteworthy because unstructured regions of proteins (that fold upon binding RNA) are a recurring theme in RNA–protein interactions (Dyson & Wright, 2002; Leulliot & Varani, 2001). The region downstream of the RNA-binding site contains a hydrophobic cluster (aa 111–143) corresponding to a short sequence reportedly homologous to Respirovirus C (Fig. 5) (Lamb & Kolakofsky, 2001). However, a motif derived from the corresponding alignment retrieves numerous unrelated, non-viral sequences (data not shown), thus reducing the reliability of the inferred relationship. In the second subgroup of Rubulavirus, the region downstream of the conserved module is widely variable and neither the Lys/Arg-rich motif nor the region of identity with Respirovirus C can be found. It is consistently predicted to be disordered by PONDR and lacks predicted secondary structure (data not shown).

Because of the shortness of the N-terminal-conserved module and of its position upstream of a disordered region, we cannot conclude whether or not it can fold alone. However, its z-helical potential and high hydrophobicity/net charge ratio indicate that it has the potential to fold in cooperation with another part of P or with another protein.

**A common modular organization in all Paramyxovirinae P**

Identification of disordered regions has allowed us to unveil a common modular organization of P. A summary of our findings on P is presented in Fig. 8. In particular, Fig. 8(A) shows the general organization of P in all Paramyxovirinae, whereas Fig. 8(B) shows the organization of P in a representative member of each of the three main genera. P has an even further modular organization than was thought previously, consisting of six distinct regions: a hydrophobic, N\textsuperscript{\textprime}a-binding region with z-helical potential, a disordered region of greatly variable length, a central region (overlapping the V ORF) than can be either ordered or disordered, a coiled-coil, a disordered linker and the XD.

**Structural flexibility: a widespread property of Mononegavirales N and P?**

The last 80 C-terminal residues of Pneumovirinae P, as well as the N-termini of Rhadoviridae and Bornaviridae P, are predicted to be in large part disordered using PONDR and HCA (data not shown). In the same vein, Filoviridae N are grossly organized into an N-terminal moiety homologous to Rhadoviridae and Paramyxoviridae N\textsubscript{CORE} (Barr et al., 1991), and a C-terminal moiety that is hypervariable, very acidic (Sanchez et al., 2001), has a low sequence complexity and contains large predicted disordered regions (data not shown). It might thus be a structural equivalent of Paramyxoviridae N\textsubscript{TAIL}. Likewise, the C-termini of Bornaviridae N are predicted to be in large part disordered using PONDR (data not shown). Taken together, these observations suggest that structural flexibility is a widespread property within the replicative complex of non-segmented, negative-stranded viruses.

**DISCUSSION**

**Disorder in the replicative complex of Paramyxovirinae**

Using complementary biocomputing methods, we have identified disordered regions of N and P and unveiled a
Fig. 4. HCA plot of SeV P. The conventions are the same as in Fig. 2. Low sequence complexity regions are underlined in light grey. Predicted or actual secondary structure elements as observed in the three-dimensional structure (Tarbouriech et al., 2000b) are shown in regular or bold style, respectively. Regions enriched in disorder-promoting residues are shaded.
Fig. 5. HCA plot of hPIV-2 P. The conventions are the same as in Figs 2 and 4. Low sequence complexity regions are underlined in light grey. Predicted secondary structure elements are shown. Note that LDRs correspond to regions poor in hydrophobic clusters, with the first and second one being strikingly rich in proline residues.
common modular organization in all Paramyxovirinae (Fig. 8). The identification of previously undetected modular regions is expected to guide functional studies. For instance, we discovered a homologous N-terminal module in the P and V proteins of Rubulavirus, Henipavirus and NDV. So far, the only identity shared by these V proteins was thought to concern their zinc-binding domain. The conserved module could be an interesting target for mutational studies aimed at elucidating the different mechanisms by which V counters interferon transduction (reviewed by Gotoh et al., 2002).

Of note is that Paramyxovirinae L contains no predicted LDR (data not shown). However, this does not exclude the presence of disordered regions shorter than the 40 aa threshold of PONDR. Indeed, the presence of a flexible hinge region in Morbillivirus L (aa 1695–1717 of MV L) has been suggested on the basis of sequence variability (McIlhatton et al., 1997). Enhanced green fluorescent protein could be inserted at this position without interfering with the function of L, which suggests that the C-terminal moiety, located downstream of the hinge, enjoys a certain degree of conformational independence (Duprex et al., 2002). This hinge region contains no low sequence complexity segments, is not visible as an interruption of hydrophobic clusters using HCA and contains predicted secondary structure elements (data not shown). This suggests that some other short flexible regions of L might escape detection using the current prediction methods.

Although no direct biochemical evidence is available, the L protein is supposed to bear most enzymatic activities required for transcription and replication, such as RNA-dependent RNA polymerase (Svenda et al., 1997) and 2′-O-methyltransferase (Ferron et al., 2002). The absence of structural disorder in the L protein might be related to the fact that a precise protein scaffold is required for these enzymatic activities.

Experimental support for disorder predictions

In most cases, experimental support exists for our predictions of disorder, although they are not always recognized as such. In retrospect, studies carried out 20 years ago on...
SeV and NDV PNT indicate that they are in large part unstructured within the viral ribonucleoprotein complex. Indeed, SeV P bound to nucleocapsids is composed of a 40 kDa C-terminal core resistant to proteolysis, while the remaining N-terminal region (extending to at least aa 221 out of 320 for PNT) is hypersensitive to proteolysis (Chinchar & Portner, 1981a; Deshpande & Portner, 1985). Likewise, NDV P is composed of a core resistant to proteolysis and an acidic region hypersensitive to proteolysis (Chinchar & Portner, 1981b). From the present knowledge of the sequence of NDV P, we can conclude that the resistant core is PCT, while it is PNT that is degraded and thus mostly disordered. Interestingly, like *Rubulavirus*, NDV PNT is composed of the conserved module followed by a predicted disordered region (data not shown). However, one still cannot reliably conclude whether the conserved module is disordered or not, since a putative 5 kDa globular peptide (the expected molecular mass of the module) might escape detection using SDS-PAGE.

Likewise, the hypersensitivity to proteolysis of *Paramyxovirinae* N

Fig. 8. Functional and structural organization of P in *Paramyxovirinae*. (A) General organization of *Paramyxovirinae* P. Globular and disordered regions are represented by large and narrow boxes, respectively. The region overlapping the V ORF is represented by dotted lines to indicate that it can be either disordered or structured (see lower panel). The line separating PNT and PCT is located as defined in Fig. 1. (B) Organization of P in the prototype member of each genus. In *Rubulavirus* and *Morbillivirus*, the central region, overlapping the V ORF, is disordered, whereas it is globular in the case of *Respirovirus*. The hydrophobic, N

supported by the protease sensitivity observed within MV (Longhi *et al.*, 2003) and SeV (Tarbouriech *et al.*, 2000a) PCT and by spectroscopic studies on SeV PCT (Marion *et al.*, 2001).

**Functional implications of structural disorder in the replicative complex**

Beyond *Paramyxovirinae*, we found that structural disorder is a widespread property of *Mononegavirales* N and P, suggesting a functional significance. In particular, since unstructured regions are considerably more extended than globular ones, the very long reach of N and P might enable them to act as linkers. Moreover, the presence of flexible regions at the surface of the viral nucleocapsid enables transient interactions with several, structurally distinct partners (Dunker *et al.*, 1998, 2001; Dunker & Obradovic, 2001; Liu *et al.*, 2002; Uversky, 2002a; Wright & Dyson, 1999). The pattern of interactions of N

Likewise, the hypersh�ibility to proteolysis of *Paramyxovirinae* N

In the same vein, the presence of a flexible linker in P, for which we could not reach the same degree of confidence than for our other predictions, is...
P are involved in manifold interactions essential for RNA transcription and replication.

**Induced folding in N and P?**

The experimental evidence mentioned above suggests that some disordered regions we describe are unstructured in vitro not only as isolated domains but also in the context of full-length proteins. SeV and NDV PNT are mostly unstructured even within P bound to nucleocapsids (Chinchar & Portner, 1981a, b; Deshpande & Portner, 1985), while the linker region of SeV P is unstructured in the context of PCT (Tarbouriech et al., 2000a). However, the possibility that these regions may fold in vivo in the presence of appropriate solute concentrations or of their physiological partner(s) (a process called ‘induced folding’) (Dyson & Wright, 2002; Uversky, 2002b) cannot be ruled out. While this manuscript was in preparation, we found (Dyson & Wright, 2002; Uversky, 2002b) cannot be ruled out. While this manuscript was in preparation, we found that MV N_TAIL undergoes such an unstructured-to-structured transition upon binding to PCT (Longhi et al., 2003). With respect to PNT, we note that the extreme N terminus of Paramyxovirinae P (especially the conserved module in Rubulavirus), which is involved in binding to N° (Curran et al., 1995b; Nishio et al., 1996; Precious et al., 1995; Shaji & Shaila, 1999; Tober et al., 1998), contains hydrophobic clusters associated with α-helical potential (Figs. 4, 5, 7 and 8). Such an α-helix could be actually induced in MV PNT in the presence of the solvent trifluoroethanol, which is used to unveil disordered regions with a propensity to undergo induced folding (Karlin et al., 2002b). Another region likely to undergo induced folding upon binding its target is the arginine-rich, RNA-binding region of Rubulavirus PNT, reminiscent of the arginine-rich motif in the disordered bacteriophage anti-termination protein N, which folds upon binding to RNA (Mogridge et al., 1998).

**Phosphorylation occurs on disordered regions of N and P**

The role of phosphorylation of Paramyxoviridae N and P is still unclear (Lamb & Kolakofsky, 2001). Remarkably, phosphorylation of SeV N occurs within N_TAIL and that of Morbillivirus and Respirivirus P occurs within PNT (Byrappa & Gupta, 1999; Byrappa et al., 1996; Das et al., 1995; Hsu & Kingsbury, 1982; Jonscher & Yates, 1997; Vidal et al., 1988). Further studies will tell whether the occurrence of phosphorylation on disordered regions of proteins is coincidental or whether it is a widespread property, as suggested by Dunker et al. (2002a). Interestingly, Zetina (2001) has recently observed that a number of intrinsically disordered proteins share a common motif, called the ‘helix-unfolding motif’, which might control the unfolding of intrinsically disordered proteins in response to cellular events, perhaps by means of phosphorylation. Although no such motif is found in Paramyxovirinae N or P, a hint that phosphorylation of N_TAIL and PNT might modulate their structural state comes from data available in the literature. Indeed, MV N° and N^NUC, which have different conformations (Gombart et al., 1995), are phosphorylated with a different pattern (Gombart et al., 1995). In the same vein, Byrappa et al. (1996) showed that all potential phosphorylation sites of SeV PNT are equally accessible to kinases, whereas once phosphorylated they have a different accessibility to phosphatases, thus suggesting that phosphorylation may affect the conformation of PNT. However, much caution is required because no study has so far been able to elucidate the tantalizing function(s) of phosphorylation of P.

**Preliminary insights from the comparison of HCA and PONDR**

The present study shows that HCA is an invaluable tool, very intuitive in qualitatively highlighting disordered regions, owing to its easy visualization of periodical, hydrophobic features directly on the primary sequence. Although its usefulness in elucidating the modular organization of proteins (which implies recognition of disordered linkers) has been proved (Callebaut et al., 1997), HCA is not widely used for the specific purpose of predicting disordered regions and is in fact not mentioned in the reviews on disorder referenced herein. We show that an HCA plot can serve as a convenient support to plot other information, such as PONDR LDRs, secondary structure predictions and low sequence complexity segments. Comparison of HCA plots and PONDR LDRs shows that, although by no means absolute, there seems to be an inverse correlation between PONDR predictions of disorder and the presence of hydrophobic clusters (Figs. 2, 4 and 5). We hope that our study will open the way to a more quantitative comparison of disorder predictors, and perhaps to their future refinement, a point crucial for current structural genomics projects.

**Disorder and lack of predicted secondary structure**

Another hallmark of disordered regions, i.e. their lack of predicted secondary structure, has not been the subject of systematic studies at the sequence level. Interestingly, however, Liu et al. (2002) recently showed the wide occurrence in proteins of long (>70 aa) regions with little or no predicted secondary structure, called NORS (for ‘no ordered regular structure’). NORS are defined as protein regions that comprise more than 70 residues with less than 12% predicted secondary structure and have at least one segment >10 aa predicted to be accessible to the solvent, as estimated by PhD (Rost, 1996). Liu et al. (2002) found that NORS overlap only partially with LDRs (i.e. some NORS are structured) and their structural significance is not well established yet. Analysis of all disordered regions >70 aa identified herein (N_TAIL, PNT and the central regions) reveals that they fulfil only the first criterion, i.e. they have less than 12% predicted secondary structure. However, the accessibility prediction values of PhD for these regions, including those the disorder of which has been biochemically proved (Karlin et al., 2002b; Longhi et al., 2003), are under the threshold of reliability of this
program. Although preliminary, this suggests that the methods of accessibility prediction might not give reliable results on disordered regions, causing some of these disordered regions to escape the classification as NORS.

Implications for structural studies of N and P

The identification of disordered regions within proteins should avoid numerous fruitless attempts to crystallize proteins (or protein domains) containing such large unstructured regions. Furthermore, some of these regions are probably unstructured even when complexed with their partner(s), thus preventing crystallization. The information on the modular organization of PCT derived from the present study led to the resolution of the crystal structure of the extreme C-terminal domain (XD) of MV P (Johansson et al., 2003), thus validating the reliability of the prediction approach described in this paper.

Once the structure of globular domains of N and P has been solved, structural advances in the study of the replicative complex will either (i) need removal of several dispensable, unstructured regions for crystallization of protein complexes, a trial-and-error process likely to be very time-consuming, (ii) rely on techniques that can deal with both ordered and disordered regions of proteins (such as small angle X-ray scattering) and (iii) rely on biocomputing methods to identify further functional regions in poorly conserved, unstructured parts of N and P.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Fondation pour la Recherche Médicale (FRM) to D. K. This study has been carried out with financial support from the Commission of the European Communities, specific RTD programme ‘Quality of Life and Management of Living Resources’, QLK2-CT2001-01225, ‘Towards Communities, specific RTD programme ‘Quality of Life and’. We wish to thank K. Dunker, J. Curran, L. Roux and D. Gerlier for useful remarks. We also thank B. Henrissat and I. Callebaut for critical advice on HCA.

REFERENCES


Protein disorder and the evolution of molecular recognition: theory, predictions and observations. Pac Symp Biocomput 473–484.


