Protein domain definition should allow for conditional disorder

Kavestri Yegambaram, Esther M.M. Bulloch, and Richard L. Kingston*

School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

Received 1 July 2013; Revised 4 August 2013; Accepted 12 August 2013

DOI: 10.1002/pro.2336

Published online 20 August 2013 proteinscience.org

Abstract: Proteins are often classified in a binary fashion as either structured or disordered. However this approach has several deficits. Firstly, protein folding is always conditional on the physiochemical environment. A protein which is structured in some circumstances will be disordered in others. Secondly, it hides a fundamental asymmetry in behavior. While all structured proteins can be unfolded through a change in environment, not all disordered proteins have the capacity for folding. Failure to accommodate these complexities confuses the definition of both protein structural domains and intrinsically disordered regions. We illustrate these points with an experimental study of a family of small binding domains, drawn from the RNA polymerase of mumps virus and its closest relatives. Assessed at face value the domains fall on a structural continuum, with folded, partially folded, and near unstructured members. Yet the disorder present in the family is conditional, and these closely related polypeptides can access the same folded state under appropriate conditions. Any heuristic definition of the protein domain emphasizing conformational stability divides this domain family in two, in a way that makes no biological sense. Structural domains would be better defined by their ability to adopt a specific tertiary structure: a structure that may or may not be realized, dependent on the circumstances. This explicitly allows for the conditional nature of protein folding, and more clearly demarcates structural domains from intrinsically disordered regions that may function without folding.

Keywords: intrinsic protein disorder; protein stability; coupled binding and folding; stabilizing osmolytes; trimethylamine N-oxide; rubulaviruses; mumps virus; RNA-dependent RNA polymerase

Introduction

For decades the field of protein science was dominated by the study of stable and folded polypeptides with well-defined structures. One recurrent feature of such proteins is that as their size increases, they generally subdivide into spatially distinct regions that can function semi-autonomously. These regions

are termed protein domains, and their continual reassortment and adaption by viruses and cells beautifully illustrates the evolutionary connections that exist throughout biology.^{1–4} Because of the repeated occurrence of similar domains in differing structural contexts, and their linkage with function, domain identification and classification underpins protein taxonomy and functional annotation.^{5–8}

Despite its importance, the definition of protein domains has never been exact, and can be approached in a number of ways. From a purely structural perspective domains are spatially distinct and compact regions of the polypeptide,⁹ whereas from a thermodynamic perspective^{10,11} the emphasis has been on the ability of domains to fold cooperatively into a stable structure. These two perspectives are naturally connected, as Wetlaufer originally speculated.¹² Domain definition is also informed by the recurrence of domains in different genetic and

Abbreviations: CD, circular dichroism; HSQC, Heteronuclear sequential quantum correlation; HPIV4b, Human Parainfluenza Virus 4b; GST, glutathione-S-transferase; MBP, maltose-binding protein; NLPs, Nucleocapsid-Like Particles; SV5, Simian Virus 5; TEV, Tobacco Etch Virus; TMAO, trimethylamine N-oxide.

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Richard L. Kingston; School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand. E-mail: rl.kingston@auckland.ac.nz

structural contexts, which is often detectable through sequence comparisons.

Although the definition of a domain remains basically heuristic, we highlight two concepts that are invoked repeatedly in standard structural and energetic approaches. The first of these is stability or foldedness, while the second is autonomy or independence. To pick just two examples, a domain is defined as a “part of a polypeptide chain that can fold independently into a stable tertiary structure” in the influential textbook by Branden and Tooze,¹³ and as “an independently structured unit capable of autonomous folding” in a recent article by Towse and Dagget.¹⁴ Innumerable variations on these definitions exist, but the concepts of stability and autonomy are recurrent. Stability implies the overwhelming population of a specific conformational and energetic state, whereas autonomy implies the independence of the folding process from any surrounding sequence in which the domain is embedded. Autonomy might also imply that folding is independent of any external influence, although that interpretation is clearly incorrect. What is highlighted in this article is that rigid adherence to many textbook definitions of the domain will arbitrarily exclude unstable proteins whose folding is strongly conditional on the physical and chemical environment, an omission that makes very little biological sense.

This is more than a minor technicality. Structure is not an invariant property of a polypeptide, unambiguously determined by amino acid sequence alone. It has long been known that most globular proteins will unfold at extremal temperatures and pressures, or in response to the presence of chemical denaturants.¹⁵ Thus the notion of protein stability has no precise meaning without specification of the physiochemical environment. More recent has been the realization that many proteins are partially or completely unstructured under “normal” biological conditions. For years these disordered proteins were simply outside of view, buried in the avalanche of high-resolution structural data being generated by X-ray crystallography. But it is now appreciated that disordered proteins are both ubiquitous, and biologically critical.¹⁶ And this creates some challenges for the definition of the protein domain.

How are “unstable” proteins that do not exist in one specific and overwhelmingly populated conformational state to be classified, and how does this merge with the original conception of the protein domain? No consensus exists. Do disordered proteins belong by themselves in their own taxonomic category?¹⁷ Are we best served by a simple binary classification of proteins as folded and stable, or disordered?¹⁸ There are several points critical to this discussion. Firstly, as already noted, a protein that is structured in some circumstances will be disordered in others. Folding is always conditional

on the physiochemical environment. Secondly, when considering disordered polypeptides of sufficient size for tertiary structure formation there is a fundamental distinction to be made. Some sequences will be strictly unfoldable, and will never adopt a specific structural state. They lack the capacity for folding. Others will partly or completely fold in response to a change in environmental conditions. We suggest that the distinction between proteins that are capable of folding and proteins that are not is fundamental, because the folded state can confer great functional specificity. For proteins that can realize a folded state, that state is likely to be integral to their function. We therefore use the terms constitutively disordered for unfoldable sequences, and conditionally disordered for foldable sequences, in order to establish this distinction.

Conditionally disordered proteins, which sit on the boundary between order and disorder,¹⁶ are what concern us here. Transition to the folded state will often be induced by specific association with other atoms or molecules (coupled folding and binding),¹⁹ or the introduction of chemical osmolytes that can non-specifically drive their folding.^{20,21} Conditionally disordered proteins can be, and generally are, classified together with their more stable counterparts on the basis of their common folded state. However this practice is inconsistent with most heuristic definitions of the protein domain, which still invoke stability as a defining property of the domain.

The need for a change in viewpoint is illustrated with an experimental study. We have surveyed the structural properties of a small domain drawn from the RNA polymerase of mumps virus and its closest relatives (the rubulaviruses). These viruses have a single-stranded RNA genome, which is organized and protected by a viral packaging protein (the N protein), creating a helical complex termed the nucleocapsid. The viral RNA polymerase must traverse the nucleocapsid during viral gene transcription and genome replication. The domains under study act as the flexibly tethered “feet” that allow the polymerase to walk along its protein–RNA template.^{22,23} The feet are located at the C-terminus of the non-catalytic subunit of the polymerase (the P protein), and must repeatedly bind and release the nucleocapsid as the polymerase translocates, requiring highly specific yet reversible interactions. For many paramyxoviruses,^{24–29} the viral subfamily that encompasses the rubulaviruses, this balance seems to be achieved by coupling binding with protein folding.¹⁹ For example, in measles virus and Sendai virus, the polymerase feet grasp a sequence within the constitutively disordered “tail” of the viral N protein, which decorates the surface of the nucleocapsid,³⁰ causing it to fold into a helix.^{24–26}

For mumps virus, the polymerase feet themselves were previously shown to be unstable.^{28,31} The 3-

Figure 1. Structure of the mumps polymerase foot domain in the folded state. The ribbon diagram shows the structure of the mumps polymerase foot domain, as determined by X-ray crystallography.²⁸ However, the compact 3-helix bundle depicted in the figure is not significantly populated in solution, where the protein exists as a dynamic ensemble of partially-formed α -helices.²⁸ The program UCSF Chimera¹⁰¹ was used to prepare this figure as well as Figure 8.

dimensional structure of the mumps foot domain was determined using X-ray crystallography, revealing the expected 3-helix bundle (Fig. 1).²⁸ However the folded state is not significantly populated in solution,²⁸ where the protein exists instead as a classical “molten globule”:^{32,33} a compact yet dynamic ensemble of partially-formed α -helices, with a nascent hydrophobic core. Complete folding of the mumps foot domain can be promoted by addition of stabilizing cosolutes such as trimethylamine *N*-oxide (TMAO).²⁸

The purpose of our current study was to establish whether the feet of all rubulavirus polymerases were similarly organized. Using an array of structural and biophysical techniques we show that the rubulavirus foot domains span the structural continuum, ranging from stable and folded to largely disordered in solution under standard conditions. However, all apparently realize a common folded state: a compact bundle of three α -helices. In many cases folding of the domain is likely to occur transiently and to be associated with binding their biological partner. This highlights the problem with using conformational stability as one of the defining properties of a protein domain.

Results

The evolutionary relationship between the studied domains, and their expected structure

The inferred phylogenetic relationship between *Rubulaviruses* with fully sequenced genomes is shown in Figure 2(A). Polymerase foot domains from Menangle Virus, Human Parainfluenza Virus 4b (HPIV4b), Simian Virus 5 (SV5), and Mapuera Virus were selected for analysis in this study, together with the previously characterized domain from mumps virus.²⁸ This

achieves coverage of all major branches of the tree, excepting the very recently identified Tuhoko viruses.³⁴ An alignment of the foot domain sequences is shown in Figure 2(B), together with a graphic representation as a sequence logo.³⁵ The mumps foot domain has between 25 and 43% amino acid sequence identity with domains in the other major branches of the tree.

All of the polymerase foot domains under study are predicted to be α -helical on the basis of their amino acid sequences alone. For example, application of the standard information theoretic method GOR V³⁶ clearly demarcates three putative helices within each sequence (Supporting Information, Fig. S1). Based on existing structural data,^{25,28,37–39} and the amphipathic character of each helix, the straightforward expectation is that these polypeptides can all fold to form a compact 3-helix bundle (Fig. 1). However the extent to which this folded state is populated in the absence of the viral nucleocapsid remains unknown. This is the point we have addressed experimentally.

Binding function

Before characterizing the structural properties of the foot domains, their ability to bind to their cognate N proteins was tested using a glutathione-*S*-transferase (GST) pull-down assay. The wild-type foot domains were fused to the C-terminus of GST, while the N proteins had a C-terminal polyhistidine tag appended (Supporting Information, Tables S1 and S3). All proteins were expressed in *Escherichia coli*. Bacterial expression of paramyxoviral N proteins invariably results in the spontaneous encapsidation of cellular RNA and the formation of Nucleocapsid-Like Particles (NLPs) of varying lengths.^{31,40–43} These NLPs provide an excellent model for the authentic viral nucleocapsid.^{44,45} Bacterial lysates containing the GST-foot domain fusion proteins and the NLPs were used as input for the pulldown assay.

The results of the assay are shown in Figure 3. The Menangle virus, HPIV4b, and SV5 foot domains specifically bind to their cognate NLPs, as was previously demonstrated for the mumps virus foot domain.³¹ For Mapuera virus, the situation remains unclear. In this case specific binding of the NLPs to the GST foot domain fusion protein does not significantly exceed the background binding observed for the negative control (GST alone). We conclude that for Mapuera virus, binding is too weak to be robustly detected using a pull-down assay. Consistent with binding of the foot domains being specific and biologically relevant, we were unable to detect binding of any foot domain to the NLPs of a different virus using the same assay procedure (data not shown).

Secondary structure content

To facilitate biophysical analysis, polymerase foot domains carrying minor sequence variations

Figure 2. Evolutionary relationships between the rubulaviruses. (A) Phylogenetic tree showing the relationship between the *Rubulaviruses*. The tree was constructed by Bayesian analysis (see Materials and Methods) and is based on the sequences of the viral L and M proteins. NCBI accession codes for the viral sequences are shown in brackets. Newcastle disease virus functions as an outgroup, and belongs to the closely related *Avulavirus* genus. Branch lengths represent relative genetic distances, while the numbers at the nodes are the estimated posterior probabilities for the existence of the clades. The horizontal bar represents 0.3 substitutions per amino acid site. (B) Alignment of *Rubulavirus* polymerase foot domain sequences and graphic representation as a sequence logo (where residues are numbered according to the mumps P sequence). Sequences were aligned using Clustal Omega.⁹⁶ Foot domain sequences studied in this article are highlighted in bold.

(Supporting Information, Table S1) were initially isolated as GST fusion proteins. The foot domains were then released from GST using specific proteolytic cleavage, and purified to chemical and physical homogeneity (see Materials and Methods).

To assess the secondary structure content of the foot domains, far UV circular dichroism (CD) spectra were recorded on the four novel foot domains, as well as the previously characterized domain from mumps virus (Fig. 4). The CD spectra demonstrate that the domains exist on a structural continuum. While all

the spectra exhibit the features expected of α -helical proteins, negative bands at 208 and 222 nm and a positive band near 192 nm, the absolute intensities vary markedly. The Menangle virus and HPIV4b foot domains are the most structured, and their CD band intensities are comparable to those of the stable and folded foot domain from measles virus^{31,46} suggesting formation of a 3-helix bundle. In contrast the mumps virus, SV5 and Mapuera virus foot domains, considered in rank order, contain progressively less α -helical structure. The mumps foot domain exists as a compact

Figure 3. The Rubulavirus polymerase foot domains are functional and interact with their cognate viral nucleocapsid proteins. The figure shows SDS-PAGE analysis of pull-down experiments testing for interactions between GST-foot domain fusion proteins and the cognate viral nucleocapsid (N) proteins. In each case, lanes 1, 2, and 3 show the inputs to the assay—bacterial lysates expressing the GST-foot fusion protein, the viral N protein, and GST. Lane 4 shows the pull-down experiment (GST-foot + N), while Lane 5 shows the negative control (GST + N). For Menangle virus, HPIV4b and SV5, specific binding of the foot domain to the viral N protein is demonstrated. For Mapuera virus, binding above background levels is not detectable using this assay.

molten globule in solution,²⁸ and the data suggested the SV5 foot domain might be similarly organized, though with even less persistent structure.

Thermostability of the domains and the effects of a stabilizing cosolute

To gain insight into the tertiary organization of the domains, temperature-induced conformational changes were monitored using the CD signal at 223 nm (Fig. 5). In essence this monitors the dissipation of helical secondary structure within the foot domains as they are heated. The effects of the stabilizing osmolyte TMAO^{20,21} on the thermal unfolding process were also examined.

For Menangle virus and HPIV4b, the unfolding curves appear characteristic of stable and folded proteins, with a sigmoidal transition region, and clear pre- and post-transition baselines. The unfolding transitions are quite broad, spanning ~40°C in each case. This is not unexpected. For a simple two-state unfolding process^{47,48} the width of the transition is

governed by the enthalpy and entropy of unfolding,^{49–51} which are themselves essentially linearly dependent on chain length.^{52,53} In other words, the width of the transition reflects the size of the cooperative unit, and small proteins, such as the polymerase foot domains, will almost certainly have broad thermal unfolding transitions, even if they adhere to the two-state model and unfold cooperatively. We note that although the two-state model⁵¹ fits the thermal unfolding data very well (not shown), this fact alone is not sufficient to establish its applicability.^{48,50}

In the presence of 0.9 M TMAO, a simple displacement of the unfolding curves to higher temperature, without significant change in the shape, is observed for both the Menangle virus and HPIV4b foot domains. This is again characteristic of the behavior of folded proteins in the presence of stabilizing osmolytes (see e.g.^{54,55}). In such cases the midpoint of the thermal unfolding transition generally exhibits a linear dependence on the osmolyte concentration.^{55,56}

Figure 4. Total α -helix content varies widely among the Rubulavirus polymerase foot domains. The figure shows far UV circular dichroism (CD) spectra of the foot domains, measured at 6°C in 5 mM $\text{NaH}_2\text{PO}_4/\text{Na}_2\text{HPO}_4$ buffer pH 7.0, 100 mM NaF. The spectra demonstrate the decreasing total α -helix content of the domains proceeding from left (most structured) to right (least structured).

Figure 5. Folding of the less structured Rubulavirus polymerase foot domains is promoted by addition of the cosolute TMAO. Thermal stability data for all five foot domains were recorded by following the CD signal at 223 nm as a function of temperature. The figure shows measurements made in both the absence (grey curves) and presence (black curves) of 0.9 M TMAO. For the stable and fully structured domains (Menangle virus and HPIV4b), addition of TMAO causes a simple rightward shift of the unfolding curves. For the unstable and partially structured domains (mumps virus, SV5 and Mapuera virus), addition of TMAO changes the appearance of the unfolding curves, which begin to take on the sigmoidal shape characteristic of folded proteins.

Overall, the thermal unfolding data (Fig. 5) are consistent with the basic CD spectra (Fig. 4), and suggest the Menangle virus and HPIV4b foot domains have folded into a stable 3-helix bundle.

Very different behavior is observed for the foot domains from mumps virus, SV5, and Mapuera virus. For these domains, none of the unfolding curves exhibit a clear pre-transition baseline. For mumps virus the helical structure begins dissipating in a nearly linear fashion, as soon as the temperature is raised from the minimum 6 °C (c.f.²⁸). The behavior of the SV5 foot domain is more complex. In this case the unfolding curve is not monotonic, and heating is associated with a small initial increase in the CD signal followed by a gradual decrease. Interestingly, the apparent cold destabilization of the partially structured SV5 foot domain closely resembles that of the fully structured ubiquitin associated domain (UBA2),⁵⁷ which forms a 3-helix bundle topologically identical to the polymerase foot. The behavior of the Mapuera virus foot domain is again straightforward, with the very small complement of α -helix dissipating both gradually and continually as the temperature is raised.

In the presence of 0.9 M TMAO striking changes are seen in the thermal unfolding curves of the three least-structured foot domains. For the mumps virus and SV5 foot domains the curves are both displaced rightward, and changed markedly in appearance, taking on the more sigmoidal shape characteristic of folded proteins. Addition of TMAO is known to promote complete folding of the mumps virus foot domain,²⁸ and we infer that the same is true for the SV5 foot domain. For the Mapuera virus foot domain the α -helical content also increases with addition of TMAO, but to a limited extent. Hints of an unfolding transition begin to emerge, which suggests a partial shift toward the folded state has been achieved.

Tertiary organization assessed using solution NMR spectroscopy

To confirm the interpretation of the CD data, we used solution NMR spectroscopy as an independent probe of the tertiary organization. 2D ¹H-¹⁵N Heteronuclear sequential quantum correlation (HSQC) spectra were collected for each of the novel foot domains (Fig. 6 top panels, Supporting Information, Fig. S2). For the partially structured SV5 and Mapuera virus foot domains HSQC spectra were also collected in the presence of 2 M TMAO (Fig. 6, bottom panels).

The HSQC experiment returns a H,N chemical shift correlation map between directly bonded nuclei. For a small and folded protein, the presence of the backbone and side chain amide groups should give rise to a set of well dispersed individual peaks of roughly equal intensity. On the other hand, a partially folded protein existing as a dynamic conformational ensemble will be characterized by more limited chemical shift dispersion. Conformational exchange on the millisecond timescale will also give rise to extreme line broadening effects and large variations in peak intensity.⁵⁸

The HSQC spectra of the Menangle virus and HPIV4b foot domains are consistent with the overwhelming population of the folded state (Supporting Information, Fig. S2), with the appropriate number of well resolved peaks present in each. In contrast, the HSQC spectra of the SV5 and Mapuera virus foot domains (Fig. 6, top panels) are consistent with the presence of a heterogeneous conformational ensemble, exhibiting limited chemical shift dispersion in the proton dimension and large variations in peak width and intensity. As inferred from the thermal unfolding studies, addition of TMAO appears to promote complete folding of the SV5 foot domain, and partial folding of the Mapuera virus foot domain.

Figure 6. NMR spectroscopy confirms that TMAO induces partial or complete folding of the unstable foot domains. The figure shows 2D ^1H - ^{15}N HSQC spectra for the SV5 and Mapuera polymerase foot domains, recorded at 25 °C in the absence (top) or presence (bottom) of 2 M TMAO. Addition of TMAO promotes complete folding of the SV5 foot domain and partial folding of the Mapuera virus foot domain, consistent with earlier observations made for mumps virus ²⁸.

(Fig. 6, bottom panels). In particular the increased chemical shift dispersion and significant line broadening observed for the Mapuera virus foot domain upon addition of TMAO⁵⁸ are consistent with transition from a near unfolded state into a molten globule. Similar effects are seen in spectra of the mumps virus and SV5 foot domains, before the TMAO-induced transition to the folded state has been completed (data not shown).

Standard Gibbs energy of unfolding for the fully structured domains

To gain further insight into the global thermodynamic stability of the polymerase foot domains, the two most structured domains under study, from Menangle virus and HPIV4b, were subject to chemical denaturation experiments using urea (Fig. 7). We interpreted the results of these experiments using the standard two-state model generally employed for quantitating protein stability.^{47,48} In this model all processes involved in folding, such as hydrophobic collapse and secondary structure formation, are assumed to occur in one

largely coincident transition, with intermediate species never significantly populated. In this case, the standard Gibbs energy associated with unfolding may be accessed experimentally. This was determined to be 13.9 kJ/mol for the Menangle virus foot domain and 13.5 kJ/mol for the HPIV4b foot domain at 15 °C. Corresponding “*m* values” for urea denaturation are −4.54 kJ/mol/M and −3.39 kJ/mol/M, respectively. The standard Gibbs energies sit at the lower extreme of those measured for small globular proteins,^{15,59} consistent with the marginal stability of the foot domains. While a useful starting point, the applicability of the two-state model still needs to be confirmed,^{48,50} particularly given the existence of homologous and functional domains which are partially disordered, and which clearly do not conform to two-state folding behavior.

3D structures determined using X-ray crystallography

In an attempt to determine the 3D structures of the previously uncharacterized polymerase foot domains,

Figure 7. Even the folded Rubulavirus polymerase foot domains are marginally stable. Urea denaturation of Menangle virus (top) and HPIV4b (bottom) polymerase foot domains was followed using the CD signal at 222 nm. The open circles represent the experimentally recorded CD data, while the solid lines show the standard two-state unfolding model that was fitted to the data. Under assumption of the two-state model, the standard Gibbs energies of unfolding (15°C) are determined to be 13.9 kJ/mol for the Menangle virus foot domain and 13.5 kJ/mol for the HPIV4b foot domain, while corresponding “m values” for urea denaturation are -4.54 kJ/mol/M and -3.39 kJ/mol/M, respectively. Approximate 95% confidence intervals for the model parameters are: Menangle $\Delta_r G^\circ$, 11.5–16.9 kJ/mol; HPIV4b $\Delta_r G^\circ$, 13.0–13.9 kJ/mol; Menangle m, -4.39 to -4.70 kJ/mol/M; HPIV4b m, -3.28 to -3.52 kJ/mol/M.

they were each fused to the C-terminus of maltose-binding protein (MBP), via a short linker sequence. For this purpose we employed variant MBP proteins (amino acids 1–367), carrying surface mutations favorable for crystallization.⁶⁰ The purified fusion proteins were subject to crystallization trials, and diffraction-quality crystals of the Menangle virus and HPIV4b foot domain fusion proteins were obtained. Following X-ray data collection, structures were determined by the method of molecular replacement. Statistics associated with X-ray data collection, model building and refinement are given in Supporting Information, Table S4.

As expected the Menangle virus foot domain forms a compact bundle of 3 α -helices, with the first helix of the foot domain extending the last helix of MBP in a semi-continuous fashion [Fig. 8(A)]. The

domain closely resembles that from mumps virus (Fig. 1), with the root-mean-square deviation for 49 backbone C α atoms equal to 0.65 Å. Less expected were the results for the HPIV4b foot domain. In this case only the first two helices of the foot domain were visualized in electron density maps, with the third helix disordered. The hydrophobic residues that line the interior face of helices 1 and 2 are buried by intermolecular association with a neighboring foot domain [Fig. 8(B)]. A homo-dimeric 4-helix bundle results, with a repacked hydrophobic core [Fig. 8(B) and Supporting Information, Fig. S3]. The same species is observed in two crystal forms, generated under different solution conditions (Supporting Information, Table S4). However, we have found no evidence that the homo-dimer persists in solution, where the spectroscopic data suggest the HPIV4b foot domain forms a monomeric 3-helix bundle (Figs. 4–6 and Supporting Information, Fig. S2). Hence even the apparently stable and folded members of this domain family are structurally malleable.

Analysis of hydrophobic core volumes for the foot domains

The volume of the hydrophobic core that can form in the interior of the polymerase foot domains is necessarily small. As hydrophobic interactions are a primary determinant of protein stability this imposes some fundamental limits on the domains. Based on simple thermodynamic arguments, Privalov⁶¹ estimated the lower bound for the emergence of cooperative folding behavior in proteins at around 50 amino acids. Excluding those proteins which oligomerize, contain extensive disulfide-bonding, or incorporate metal-binding sites, the smallest natural and stable proteins identified experimentally are around 40 amino acids in length (see e.g.^{62,63}). While shorter synthetic peptides with stable structures do exist (see e.g.^{64,65}) it seems clear that limited chain length restricts protein stability, in part due to the difficulty of sequestering hydrophobic groups from water.^{66–68}

To explore this idea we estimated the hydrophobic core volumes for the polymerase foot domains, using a weighted Voronoi tessellation procedure (see Materials and Methods). The volume associated with the non-polar atoms buried in the protein interior is estimated as 1127 Å³ for the Menangle polymerase foot domain, and 952 Å³ for the mumps polymerase foot domain. These values approximate the minimum stable hydrophobic core size of ~ 1000 Å³ suggested by Tsai, Gerstein and Levitt,⁶⁸ based on their simulations of the aggregation of non-polar solutes in water. For comparison we estimated the hydrophobic core volume of several other small and stable proteins (34–53 amino acids) using the same procedure (Supporting Information, Table S5). These volumes ranged from 832 Å³ for the 34 amino acid

Figure 8. Foot domain structure analyzed using X-ray crystallography. The foot domains from Menangle virus and HPIV4b were fused to MBP, and the structure of the fusion proteins determined using X-ray crystallography. The structures are shown in ribbon representation, with the MBP moieties made near transparent for clarity. For Menangle virus (left) the expected 3-helix bundle was visualized. For HPIV4b (right) a variant structure was observed. The third helix of the HPIV4b foot domain is disordered and not visualized in the crystal structure. The otherwise exposed hydrophobic residues in the interior of the foot domain are buried by inter-molecular association with a neighboring foot domain, creating a homo-dimeric 4-helix bundle. Essentially the same structure was observed in two different crystal forms; however, only one of these is displayed. Detail of the bundle is displayed in Supporting Information, Fig. S3.

villin headpiece subdomain,⁶⁹ to 1025 Å³ for the 45 amino acid ubiquitin-associated domain of Fas-associated factor 1⁷⁰. Based on this analysis, the polymerase foot domains are operating close to the lower size limit for foldable sequences. We suggest that this allows for the ready emergence of highly dynamic behavior within this domain family, in which structural fluctuations become so extensive that description in terms of a single conformational state is no longer appropriate.

Discussion

The experiments described in this article explored the structural characteristics of a family of small binding domains drawn from the polymerase of mumps virus and its closest relatives (Fig. 2). These domains mediate attachment of the polymerase to its template, the viral nucleocapsid, during RNA synthesis (Fig. 3). Based on sequence alone, it would be expected that these very similar polypeptides form a compact bundle of 3 α-helices (Figs. 1 and Supporting Information, Fig. S1). The experimental reality is different, revealing a striking diversity in structural content (Fig. 4). While several of the domains (Menangle virus, HPIV4b) do adopt a specific fold, several more (mumps virus, SV5) exist natively as molten globules, and one (Mapuera virus) contains very little persistent structure at all. The less structured domains can be “force folded”, in part or in full, through addition of the generic stabilizing osmolyte TMAO (Figs. 5 and 6, see also²⁸). It seems probable that in a biological setting the rubulavirus polymerase foot domains bind to their cognate nucleocapsid in an essentially

conserved manner, but in some instances folding of the foot domain is realized only transiently during the binding cycle. While the structurally heterogeneous foot domains could be classified as intrinsically disordered, doing so does not reflect the central role of the folded state in their biological function. Hence we have termed them conditionally disordered, emphasizing their ability to realize the folded state under appropriate conditions.

For this family of domains, the linkages between structure formation and the physiochemical environment cannot be readily ignored. Some members of the family are stable under standard conditions, with essentially complete population of a unique folded state. Others are unstable under the same conditions, lacking structural and energetic specificity, and existing in an ensemble of interconverting states. The previously characterized mumps foot domain²⁸ is now seen to sit in the middle of a structural continuum (Figs. 4 and 5), flanked by domains with both more and less structure. For the fully structured domains from Menangle virus and HPIV4b, folding may be an essentially two-state process. For both of these domains the standard Gibbs energy of unfolding,⁷¹ which is the usual metric of protein stability, was experimentally determined to be ~14 kJ/mol (Fig. 7). This value is small in comparison with other folded proteins. The marginal stability of the entire domain family was reinforced by the crystallographic results. While the expected 3-helix bundle was visualized for Menangle virus, a variant structure was visualized for HPIV4b, in which displacement of the final helix and partial repacking of the hydrophobic core resulted from a

crystal packing interaction [Fig. 8(B) and Supporting Information, Fig. S3].

When describing the structural characteristics of these domains, some care is needed in the specification of the environmental conditions. For example, our assessment of secondary structure content of the foot domains (Fig. 4) was performed at 6 °C. Elevation of temperature to the host physiological range (35–40 °C) where the viral replication machinery actually functions, significantly perturbs the conformational ensemble in which the mumps virus and SV5 foot domains exist (Fig. 5). The behavior of the SV5 foot domain as a function of temperature (Fig. 5) demonstrates that heating does not always result in the simple dissipation of structure.

The impact of the highly crowded cellular environment on the structural properties of the foot domains must also be considered. All of our observations on the foot domains have been made *in vitro*, in relatively dilute solution. Excluded volume theory⁷² predicts that macromolecular crowding will heavily favor compact folded states over expanded states. However, when the effects of macromolecular crowding on protein stability have been evaluated experimentally, they have often been found to be negligible, most probably because the contribution of non-specific chemical interactions with the crowding agent can readily overwhelm simple volume exclusion effects.⁷³ The potential for non-specific chemical interactions to modulate foot domain structure is evident in the response of the domains to inclusion of TMAO (Figs. 5 and 6). Hence, while it is undoubtedly the case that the stability of the foot domains may be “tuned” within the cellular environment, understanding such effects requires detailed experimental investigation.

The structural and energetic properties of the rubulavirus polymerase foot domains likely originate with their small size (50 amino acids). For such small domains, conformational fluctuations are likely to become extremely significant,⁷⁴ and structural and energetic specificity are readily lost. Indeed analysis of hydrophobic core volumes (Supporting Information, Table S5) suggests the domains are operating close to the lower size limit for proteins stabilized by the burial of hydrophobic amino acids. Additionally, as the biological function of these domains is to mediate a reversible binding process, there may be little selective pressure to maintain the domains in the folded state—indeed the reverse could be true.

However, there is no reason to think our observations represent an isolated or uncommon scenario. Coupled folding and binding events are deeply embedded in biology, and while most of these involve the structuring of relatively short peptides, some cases involving the structuring of entire protein domains have been identified. A particularly well characterized example is the protein subunit of *Bacillus subtilis* RNase P, which comprises a single

domain of around 120 amino acids with a mixed α/β structure.⁷⁵ This protein conditionally disorders when isolated from RNA and stabilizing anions,⁷⁶ while its folding can be induced by TMAO.⁷⁷ A second well studied example is Human Lymphoid enhancer-binding factor 1 (LEF-1), a transcriptional activator of around 90 amino acids with an all α -structure. LEF-1 conditionally disorders in the absence of bound DNA.⁷⁸ In both cases, homologous proteins are known which do not exhibit the same conditional unfolding behavior. Hence similar to the situation studied here, the taxonomic result is a family of related domains, some members of which are disordered under “normal” conditions.

Because X-ray crystallography remains the predominant method of protein structure determination, there are several reasons to think conditional disorder is more significant and more widely distributed than currently recognized. Firstly, a selection bias exists—conditionally disordered proteins will be more difficult to crystallize. If follow up investigations are carried out at all, they may conclude that the protein is intrinsically and constitutively disordered. Secondly, an observation bias exists—if conditionally disordered proteins are crystallized in the presence of the ions or molecules that promote their folding, the phenomenon will go undetected until appropriate solution studies are performed.

At present experimental investigation provides the only reliable method for identification of conditionally disordered proteins. Naturally enough, the structurally heterogeneous foot domains have none of the identifiers routinely associated with intrinsically disordered sequences: they are not characterized by amino acid compositional bias or low sequence complexity (Fig. 2). Stabilizing osmolytes, such as TMAO, are particularly valuable for the investigation of such proteins, as their action is non-specific and universal. The effects of TMAO on conditionally disordered proteins can be dramatic (Figs. 5 and 6) and highly diagnostic of the potential for structure formation in apparently disordered sequences. In our case, and other instances where the unstable protein participates in a coupled folding and binding event,⁷⁷ TMAO allows for the effective experimental decoupling of these two processes.

The phenomenon of protein metamorphism provides a slightly different context in which conditional folding behavior cannot be ignored. Metamorphic proteins are capable of existing in two or more structurally specific states. The population of each state is frequently dependent on the chemical and physical environment. The human chemokine Lymphotactin (Ltn) provides an example. This 92 amino acid protein can exist in one of two specific structural states; one a monomeric mixed α/β structure, and the other a dimeric all β variant with a totally different hydrogen bond network. The proportions of each state change

with salinity and temperature.⁷⁹ This reinforces the point that protein structure can be context dependent, even in physiological buffers at regular temperature and pressure.

If the links between structure formation and the physiochemical environment cannot be conveniently set aside, what are the implications for protein domain definition? By any sensible criteria the proteins described in this article constitute a family of domains, related by their evolutionary history, sequence, structure, and function. We have used the term domain without restraint. Yet the simple fact is that many heuristic definitions of the protein domain embedded in the literature, with their emphasis on global thermodynamic and conformational stability, would specifically exclude this grouping. Hence, the language and intent of those definitions needs revisiting.

We suggest that a heuristic definition of the protein domain should focus on the ability of a polypeptide to fold, and not the exclusive population of the folded state. In this fashion, a domain can be defined as a polypeptide that can adopt a specific tertiary structure under appropriate conditions, and whose structural and energetic properties are unaffected by excision from any surrounding sequence. This definition removes the emphasis on domain stability, which cannot in any case be meaningfully defined without reference to the physiochemical environment, and clarifies the meaning of autonomy. It assimilates conditionally disordered domains such as those described in this article. The emphasis on tertiary structure implies the formation of non-local contacts between amino acids, and excludes linear peptidic motifs. These frequently become structured upon binding to a larger macromolecule, but are not capable of structure formation of their own accord. As such, they fall outside most accepted definitions of the protein domain.

While the problems we have highlighted are not new,¹⁷ in this article we have suggested how the emphasis in domain definition might be subtly shifted, to remove the troubling inconsistency that arises at the boundary between order and disorder. The potential for folding, and the realization of the folded state are not the same thing, and the emphasis in domain definition should be on the former. Some would probably go further. Studies of intrinsically disordered proteins have de-emphasized the role of folding in protein function. In particular, the identification of constitutively disordered proteins which do not fold in order to function (see e.g.⁸⁰), has led some to consider that domains are best defined on the basis of their recurrence in differing genetic and structural contexts, without regard to their structural state. Whether the term domain should be extended to encompass polypeptides which cannot realize a specific tertiary structure will be decided in the literature. We suggest the distinction between polypeptides that can and can't fold is fun-

damental, and the terms conditionally and constitutively disordered may help make this distinction.

Regardless, the way domains are defined needs to shift, if the concept of the domain is to maintain coherence when applied to conditionally disordered proteins. Domain definition should not emphasize conformational stability, and should instead focus on the capacity of the protein to attain a specific conformational state. While the attainment of a folded state can confer biological specificity, biological function could require that this state have a short mean lifetime or sparse equilibrium population. This appears particularly likely for binding proteins. For a long time domain definition has ignored the linkage between structure and the physiochemical environment, because many proteins fold into a singular and structurally specific state at regular temperature and pressure, and in the absence of other molecules. The discovery of increasing numbers of proteins that do not behave in this fashion requires a changed emphasis in heuristic definitions of the protein domain.

Materials and Methods

Construction of expression plasmids

Sequences encoding both the polymerase foot domains and the full length N proteins of HPIV4b (Genbank Accession AB543337), Mapuera virus (Genbank Accession NC_009489), Menangle virus (Genbank Accession NC_007620), and SV5 (Genbank Accession NC_006430) were synthesized by GeneArt AG. All nucleotide sequences were codon optimized for efficient heterologous expression in *Escherichia coli*. The synthetic genes were used to construct three sets of expression plasmids. The first set of plasmids expressed the foot domains fused to the C-terminus of GST (Supporting Information, Table S1). An inter-leaving Tobacco Etch Virus (TEV) protease recognition sequence allowed release of the foot domains. The second set of plasmids expressed the foot domains fused to the C-terminus of MBP (Supporting Information, Table S2). The third set of plasmids expressed the N proteins with a C-terminal polyhistidine tag appended (Supporting Information, Table S3).

For the crystallographic and biophysical analysis, several modifications were made to the foot domain sequences. A surface-exposed cysteine residue at the end of helix 1 was mutated to serine to limit chemical reactivity (c.f.²⁸). Additionally, a non-native tyrosine was appended to the C-terminus of some foot domains to facilitate protein quantitation. The mumps virus foot domain employed in some experiments was produced using a previously described plasmid.²⁸

Heterologous protein expression and protein purification

All proteins were expressed in *E. coli* BL21(DE3) (Stratagene). Transformed bacteria were cultured in

LB media supplemented with 1%(v/v) glycerol and appropriate antibiotics. Cultures were shaken in flasks at 37 °C until the optical density at 600 nm was 0.8–1.0. Protein expression was then induced by addition of isopropylthiogalactopyranoside to a final concentration of 0.5 mM, and cultures maintained at 18–28 °C for 5–18 hours, before harvesting by centrifugation. Uniform isotopic labeling of proteins for solution NMR studies was achieved by growing the bacteria in minimal media⁸¹ with ¹⁵NH₄Cl as the sole nitrogen source.

To purify the polymerase foot domains, bacteria expressing the GST fusion proteins were resuspended in lysis buffer. This was either 12.5 mM MOPS/KOH pH 7.0, 150 mM NaCl (mumps virus, Menangle virus); 12.5 mM Tris/HCl pH 8.5, 150 mM NaCl (SV5, HPIV4b); or 12.5 mM Tris/HCl pH 8.5, 150 mM NaCl, 0.5 mM EDTA, 10%(v/v) glycerol (Mapuera virus). A protease inhibitor cocktail (Complete Mini; EDTA-free, Roche Applied Science) was added, and the bacteria were lysed by high pressure shearing in a commercial processor (Microfluidics). The lysate was clarified by low-speed centrifugation and the GST fusion proteins isolated using a bed of glutathione agarose resin (Pierce). The completely disordered Mapuera foot domain generally co-purified with the *E. coli* chaperone DnaK, which was removed with an ATP wash step. The polymerase foot domains were released from GST by incubating the resin-bound fusion proteins overnight with TEV protease, supplementing the buffer with 2 mM DTT. The foot domains were eluted from the affinity resin and, excepting the Mapuera virus foot domain, further purified by cation-exchange chromatography. Proteins were bound to SP Sepharose HP resin (Amersham Biosciences) buffered with 12.5 mM malic acid/KOH pH 3.7, 50 mM NaCl, and then displaced with a linear salt gradient. The Mapuera virus foot domain was purified using an alternative procedure, due to its propensity to bind irreversibly to chromatographic resins. Firstly it was dialyzed into 5 mM NaH₂PO₄/Na₂HPO₄ pH 7.0, 100 mM NaF. Subsequently the protein solution was heated at 70 °C for 20 minutes and cooled on ice for 10 minutes. High speed centrifugation removed precipitated proteins, leaving the sample pure.

To purify the MBP-foot domain fusion proteins, bacteria expressing the fusion proteins were resuspended in lysis buffer: 12.5 mM Tris/HCl pH 8.5, 150 mM NaCl, 0.5 mM EDTA. Bacterial lysis and clarification was performed as described above. The MBP fusion proteins were isolated using a bed of amylose resin (New England Biolabs), and eluted in lysis buffer supplemented with 30 mM maltose. Proteins were further purified using anion exchange chromatography, binding them to Q Sepharose HP resin (Amersham Biosciences) buffered with 12.5 mM Tris/HCl pH 8.5, 50 mM NaCl, and then displacing them with a linear salt gradient.

Protein purity was assessed using SDS-PAGE, coupled with colloidal Coomassie staining for protein detection. Masses of all foot domains were confirmed using MALDI-TOF mass spectrometry. The concentrations of all purified proteins were determined using absorbance measurements at 280 nm.⁸²

GST pulldown experiments

Pulldown experiments were carried out in Mini Spin columns (Epoch Life Sciences) at ambient temperature. Columns were prepared by adding 100 µL of glutathione agarose resin (Pierce), equilibrated with either buffer A (12.5 mM MOPS/KOH pH 7.0, 250 mM NaCl) or Buffer B (12.5 mM MOPS/KOH pH 7.0, 50 mM NaCl, 100 mM KSCN). Buffer B was used when non-specific protein interactions with the resin were observed.⁸³ Bacterial cell pellets (10 mL cultures) were re-suspended in the appropriate buffer, lysed by sonication, and clarified by low speed centrifugation. Lysate (800 µL) containing the GST-foot fusion protein was applied to the columns, incubated for 5 minutes and then washed three times with 600 µL of buffer. Lysate (800 µL) containing the respective N proteins was then applied to the resin, incubated and washed in the same fashion. Negative controls were performed by substituting the GST fusion protein with GST alone. Proteins bound to the resin following the final wash step were analyzed directly using SDS-PAGE.

CD spectroscopy

All CD measurements were made on a PiStar spectrometer (Applied Photophysics). Protein spectra and thermal unfolding data are reported in terms of the mean residue ellipticity (θ) (deg.cm²/dmol), calculated as follows:

$$\theta = S / (10 \times c \times L \times n) \quad (1)$$

where S is the raw CD signal in millidegrees, c is the protein concentration (M), L is the cuvette path length (cm), and n is the number of peptide bonds in the protein.

Far UV CD spectra (180–300 nm) were collected on proteins at concentrations of 2–10 µM, dialyzed into 5 mM NaH₂PO₄/Na₂HPO₄ pH 7.0, 100 mM NaF. Measurements were made using a 1 mm path length quartz cuvette (Hellma Analytics) at a temperature of 6 °C, with a 1–2 nm optical bandwidth. Baseline spectra were collected on buffer alone; smoothed;⁸⁴ and then subtracted from the raw protein spectra.

Thermal unfolding data were collected on proteins at 3–10 µM concentration, dialyzed into 5 mM NaH₂PO₄/Na₂HPO₄ pH 7.0, 50 mM NaCl ± 0.9 M TMAO. TMAO concentration was determined by refractive index measurement at 589 nm.⁸⁵ CD measurements were made using a 5 mm path length quartz cuvette (Starna), with continual stirring of the sample.

Protein unfolding was followed by monitoring the CD signal at 223 nm (2 nm bandwidth), while applying a constant temperature ramp (1°C/min between 6 and 94 °C). The actual sample temperature was measured with an in-cell probe. To assess thermodynamic reversibility, repeat heating and cooling cycles were performed with each sample. In the absence of TMAO, thermal unfolding of the foot domains was fully reversible. In the presence of TMAO, thermal unfolding of some foot domains was partially reversible, most likely due to protein aggregation at elevated temperatures.

For chemical unfolding measurements, proteins were dialyzed into 20 mM NaH₂PO₄/Na₂HPO₄ pH 7.0, 50 mM NaCl. A 10 M stock urea solution was prepared in the same buffer, with the absolute urea concentration determined by both refractive index

and density measurements.⁸⁶ Prior to data collection, standard reversibility tests were performed.⁵⁶ By mixing the stock urea and protein solutions, a dilution series was prepared with a fixed protein concentration of 3 μM, and a urea concentration of 0–8 M. Each solution was transferred into a 5 mm path length quartz cuvette (Starna) and incubated at 15 °C for 1 hour before making spectroscopic measurements. A far UV CD spectrum was collected, followed by continuous measurements at 222 nm (2 nm bandwidth). Data were fitted to a two-state unfolding model, assuming the standard Gibbs energy of unfolding to be linearly dependent on denaturant concentration. In this case, the average CD measurements at 222 nm, reported in differential absorbance ΔA, are described by:^{51,56,87}

$$\Delta A = [P_T]L \left\{ \frac{(\Delta \epsilon_{F,m} + \beta_F(c - c_m)) + (\Delta \epsilon_{U,m} + \beta_U(c - c_m)) \exp(-m(c - c_m)/RT)}{\exp(-m(c - c_m)/RT) + 1} \right\} \quad (2)$$

where [P_T] is the total protein concentration (M); L is the optical path length (cm); Δε_{F,m} and Δε_{U,m} are the molar CD extinction coefficients of the folded and unfolded protein at the denaturation midpoint (M⁻¹cm⁻¹); β_F and β_U express the assumed linear dependence of the extinction coefficients on denaturant concentration (M⁻¹cm⁻¹/M); *m* expresses the assumed linear dependence of the Gibbs energy of unfolding on denaturant concentration (J mol⁻¹/M); *c* is the denaturant concentration (M); *c*_m is the denaturation midpoint (M); R is the Gas constant (J K⁻¹mol⁻¹); and *T* is the temperature (K).

Subsequently, the standard Gibbs energy of unfolding, Δ_rG[°] at any denaturant concentration (*c*) can be calculated using:

$$\Delta_r G^\circ(c) = m(c - c_m) \quad (3)$$

Model parameters were determined by non-linear least squares fit of Eq. (2) to the data. Confidence intervals for the fit parameters were approximated using a grid search method.⁸⁸

NMR spectroscopy

For solution NMR spectroscopy, uniformly ¹⁵N-labeled foot domains were transferred into 10 mM MES/KOH pH 6.0, 50 mM NaCl, supplemented with a small amount of 4,4-dimethyl-4-silapentane-1-sulfonic acid as a chemical shift reference. 2D HSQC spectra were recorded for each sample. For the less structured SV5 and Mapuera virus foot domains, spectra were also recorded in the presence of the stabilizing osmolyte TMAO. In these cases the protein sample was split in two prior to data collection. Solid deuterated TMAO

was dissolved in the first part to a final concentration of 2 M, while appropriate dilutions were performed on the second part so that protein concentrations remained equivalent. ²H₂O [final concentration 7% (v/v)] was added to all samples before NMR data acquisition. Final protein concentrations were 660 μM (Menangle virus), 500 μM (SV5), 430 μM (HPIV4b), and 10 μM (Mapuera virus), the latter domain being sparingly soluble. Spectra were recorded at a temperature of 25 °C using a 600 MHz Bruker spectrometer, equipped with a triple-resonance pulsed-field gradient cryoprobe. Data were processed using the program TopSpin [Bruker corp.].

X-ray crystallography

For crystallization, MBP-foot domain fusion proteins (concentration 150–1600 μM) were transferred into 12.5 mM Tris/HCl pH 8.5, 150 mM NaCl, 5 mM maltose. Protein crystals were grown using the sitting-drop vapor diffusion method, with crystallization conditions identified using orthogonal array-based screening experiments⁸⁹ (Details in Supporting Information, Table S4). Prior to data collection, crystals were transferred into cryo-protective buffers, and vitrified by direct immersion in liquid nitrogen. Diffraction data were collected by the oscillation method, using both laboratory and synchrotron radiation sources, with crystals maintained at 100–110 K in a cold gas stream throughout. Data integration and scaling were performed with the program HKL2000⁹⁰. All structures were determined by the method of molecular replacement with the program Phaser,⁹¹ using MBP (PDB accession code 1HSJ) as the search model. Iterative model building and refinement were carried

out with the programs COOT⁹² and Refmac.⁹³ Statistics associated with the X-ray diffraction data and the atomic models are given in Supporting Information, Table S4.

Phylogenetic analysis

The phylogenetic analysis was performed using the amino acid sequences of the *Rubulavirus* L and M proteins, as these are the most conserved among the major structural proteins of the paramyxoviruses.^{94,95} The L and M sequences were retrieved from NCBI Genbank, concatenated, and aligned using Clustal Omega.⁹⁶ Sequences from Newcastle disease virus, the prototypical member of the closely related *Avulavirus* genus, were incorporated as an outgroup to allow rooting of the tree. Consistent phylogenetic relationships were observed independent of the tree building method or evolutionary model employed. The tree presented in Fig. 1(A) was constructed with the program MrBayes,⁹⁷ assuming the Whelan and Goldman evolutionary model⁹⁸ and using a gamma distribution to describe the rate of variation across sites. Phylogenetic trees were edited and displayed using the program FigTree (<http://tree.bio.ed.ac.uk>).

Voronoi calculations

Beginning with an atomic model, additively weighted Voronoi cells were computed for each atom in the model using the program Voronoia.⁹⁹ Following Tsai, Gerstein and Levitt,⁶⁸ the hydrophobic core volume was approximated as the summed Voronoi cell volumes of all completely buried carbon atoms (i.e. those which make no contribution to the molecular surface). Weighted Voronoi tessellations were inspected visually using the program Voroprot.¹⁰⁰

Accession codes

The crystallographic models and the diffraction data have been deposited with the PDB (Menangle virus, accession code 4KYC; HPIV4b, accession codes 4KYD and 4KYE).

Acknowledgments

We thank Felix Hart for constructing some of the expression plasmids used in this study and Dr Howard Ross (University of Auckland) for advice on phylogenetic analysis. KY was supported by a University of Auckland doctoral scholarship.

References

1. Abrescia NGA, Bamford DH, Grimes JM, Stuart DI (2012) Structure unifies the viral universe. *Annu Rev Biochem* 81:795–822.
2. Levitt M (2009) Nature of the protein universe. *Proc Natl Acad Sci USA* 106:11079–11084.
3. Moore AD, Björklund AK, Ekman D, Bornberg-Bauer E, Elofsson A (2008) Arrangements in the modular evolution of proteins. *Trends Biochem Sci* 33:444–451.
4. Ponting CP, Russell RR (2002) The natural history of protein domains. *Annu Rev Biophys Biomol Struct* 31:45–71.
5. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, McMenamin C, Mi H, Mutowo-Muellenet P, Mulder N, Natale D, Orengo C, Pesceat S, Punta M, Quinn AF, Rivoire C, Sangrador-Vegas A, Selengut JD, Sigrist CJA, Scheremetjew M, Tate J, Thimmajananathan M, Thomas PD, Wu CH, Yeats C, Yong S-Y. (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res* 40:D306–D312.
6. Bru C, Courcelle E, Carrère S, Beausse Y, Dalmar S, Kahn D (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res* 33:D212–D215.
7. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH – a hierarchic classification of protein domain structures. *Structure* 5:1093–1108.
8. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536–540.
9. Richardson JS (1981) The anatomy and taxonomy of protein structure. *Adv Protein Chem* 34:167–339.
10. Privalov PL (1982) Stability of proteins: Proteins which do not present a single cooperative system. *Adv Protein Chem* 35:1–104.
11. Porter LL, Rose GD (2012) A thermodynamic definition of protein domains. *Proc Natl Acad Sci USA* 109:9420–9425.
12. Wetlaufer DB (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci USA* 70:697–701.
13. Branden C, Tooze J (1999) Introduction to protein structure. 2nd ed. Garland Publishing, New York.
14. Towse CL, Daggett V (2012) When a domain is not a domain, and why it is important to properly filter proteins in databases: conflicting definitions and fold classification systems for structural domains make filtering of such databases imperative. *BioEssays* 34:1060–1069.
15. Privalov PL (1979) Stability of proteins: small globular proteins. *Adv Protein Chem* 33:167–241.
16. Babu MM, Kriwacki RW, Pappu RV (2012) Structural biology. Versatility from protein disorder. *Science* 337:1460–1461.
17. Tompa P, Fuxreiter M, Oldfield CJ, Simon I, Dunker AK, Uversky VN (2009) Close encounters of the third kind: disordered domains and the interactions of proteins. *BioEssays* 31:328–335.
18. Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B (2011) Protein disorder—a breakthrough invention of evolution? *Curr Opin Struct Biol* 21:412–418.
19. Wright PE, Dyson HJ (2009) Linking folding and binding. *Curr Opin Struct Biol* 19:31–38.
20. Street TO, Bolen DW, Rose GD (2006) A molecular mechanism for osmolyte-induced protein stability. *Proc Natl Acad Sci USA* 103:13997–14002.
21. Baskakov I, Bolen DW (1998) Forcing thermodynamically unfolded proteins to fold. *J Biol Chem* 273:4831–4834.

22. Whelan SPJ, Barr JN, Wertz GW (2004) Transcription and replication of nonsegmented negative-strand RNA viruses. *Curr Top Microbiol Immunol* 283:61–119.
23. Kolakofsky D, Le Mercier P, Iseni F, Garcin D (2004) Viral RNA polymerase scanning and the gymnastics of Sendai virus RNA synthesis. *Virology* 318:463–473.
24. Longhi S, Receveur-Bréchet V, Karlin D, Johansson K, Darbon H, Bhella D, Yeo R, Finet S, Canard B (2003) The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. *J Biol Chem* 278:18638–18648.
25. Kingston RL, Hamel DJ, Gay LS, Dahlquist FW, Matthews BW (2004) Structural basis for the attachment of a paramyxoviral polymerase to its template. *Proc Natl Acad Sci USA* 101:8301–8306.
26. Houben K, Marion D, Tarbouriech N, Ruigrok RWH, Blanchard L (2007) Interaction of the C-terminal domains of sendai virus N and P proteins: comparison of polymerase-nucleocapsid interactions within the paramyxovirus family. *J Virol* 81:6807–6816.
27. Jensen MR, Houben K, Lescop E, Blanchard L, Ruigrok RWH, Blackledge M (2008) Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of Sendai virus nucleoprotein. *J Am Chem Soc* 130:8055–8061.
28. Kingston RL, Gay LS, Baase WS, Matthews BW (2008) Structure of the nucleocapsid-binding domain from the mumps virus polymerase; an example of protein folding induced by crystallization. *J Mol Biol* 379:719–731.
29. Habchi J, Blangy S, Mamelli L, Jensen MR, Blackledge M, Darbon H, Oglesbee M, Shu Y, Longhi S (2011) Characterization of the interactions between the nucleoprotein and the phosphoprotein of Henipavirus. *J Biol Chem* 286:13583–13602.
30. Jensen MR, Communie G, Ribeiro EA, Martinez N, Desfosses A, Salmon L, Mollica L, Gabel F, Jamin M, Longhi S, Ruigrok RWH, Blackledge M. (2011) Intrinsic disorder in measles virus nucleocapsids. *Proc Natl Acad Sci USA* 108:9839–9844.
31. Kingston RL, Baase WA, Gay LS (2004) Characterization of nucleocapsid binding by the measles virus and mumps virus phosphoproteins. *J Virol* 78:8630–8640.
32. Freire E (1995) Thermodynamics of partly folded intermediates in proteins. *Annu Rev Biophys Biomol Struct* 24:141–165.
33. Ptitsyn OB (1995) How the molten globule became. *Trends Biochem Sci* 20:376–379.
34. Lau SKP, Woo PCY, Wong BHL, Wong AYP, Tsoi H-W, Wang M, Lee P, Xu H, Poon RWS, Guo R, Li KSM, Chan K-H, Zheng B-J, Yuen K-Y. (2010) Identification and complete genome analysis of three novel paramyxoviruses, Tuhoko virus 1, 2 and 3, in fruit bats from China. *Virology* 404:106–116.
35. Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 18:6097–6100.
36. Kloczkowski A, Ting K-L, Jernigan RL, Garnier J (2002) Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *Proteins* 49:154–166.
37. Johansson K, Bourhis J-M, Campanacci V, Cambillau C, Canard B, Longhi S (2003) Crystal structure of the measles virus phosphoprotein domain responsible for the induced folding of the C-terminal domain of the nucleoprotein. *J Biol Chem* 278:44567–44573.
38. Blanchard L, Tarbouriech N, Blackledge M, Timmins P, Burmeister WP, Ruigrok RWH, Marion D (2004) Structure and dynamics of the nucleocapsid-binding domain of the Sendai virus phosphoprotein in solution. *Virology* 319:201–211.
39. Houben K, Blanchard L, Blackledge M, Marion D (2007) Intrinsic dynamics of the partly unstructured PX domain from the Sendai virus RNA polymerase cofactor P. *Biophys J* 93:2830–2844.
40. Warnes A, Fooks AR, Dowsett AB, Wilkinson GW, Stephenson JR (1995) Expression of the measles virus nucleoprotein gene in *Escherichia coli* and assembly of nucleocapsid-like structures. *Gene* 160:173–178.
41. Karlin D, Longhi S, Canard B (2002) Substitution of two residues in the measles virus nucleoprotein results in an impaired self-association. *Virology* 302:420–432.
42. Tan WS, Ong ST, Eshaghi M, Foo S-S, Yusoff K (2004) Solubility, immunogenicity and physical properties of the nucleocapsid protein of Nipah virus produced in *Escherichia coli*. *J Med Virol* 73:105–112.
43. Cox R, Green TJ, Qiu S, Kang J, Tsao J, Prevelige PE, He B, Luo M (2009) Characterization of a mumps virus nucleocapsidlike particle. *J Virol* 83:11402–11406.
44. Bhella D, Ralph A, Yeo RP (2004) Conformational flexibility in recombinant measles virus nucleocapsids visualised by cryo-negative stain electron microscopy and real-space helical reconstruction. *J Mol Biol* 340:319–331.
45. Schoehn G, Mavrikakis M, Albertini A, Wade R, Hoenger A, Ruigrok RWH (2004) The 12 Å structure of trypsin-treated measles virus N-RNA. *J Mol Biol* 339:301–312.
46. Shu Y, Habchi J, Costanzo S, Padilla A, Brunel J, Gerlier D, Oglesbee M, Longhi S (2012) Plasticity in structural and functional interactions between the phosphoprotein and nucleoprotein of measles virus. *J Biol Chem* 287:11951–11967.
47. Barrick D (2009) The studies of two-state folders and the unanswered questions about two-state protein folding. *Phys Biol* 6:015001.
48. Lumry R, Biltonen R (1966) Validity of the “two-state” hypothesis for conformational transitions of proteins. *Biopolymers* 4:917–944.
49. Schellman JA (1987) The thermodynamic stability of proteins. *Annu Rev Biophys Chem* 16:115–137.
50. Eftink MR, Ionescu R (1997) Thermodynamics of protein unfolding: questions pertinent to testing the validity of the two-state model. *Biophys Chem* 64:175–197.
51. Martin SR, Schilstra MJ (2008) Circular dichroism and its application to the study of biomolecules. *Methods Cell Biol* 84:263–293.
52. Robertson A, Murphy K (1997) Protein structure and the energetics of protein stability. *Chem Rev* 97:1251–1268.
53. Ghosh K, Dill KA (2009) Computing protein stabilities from their chain lengths. *Proc Natl Acad Sci USA* 106:10649–10654.
54. Singh R, Haque I, Ahmad F (2005) Counteracting osmolyte trimethylamine N-oxide destabilizes proteins at pH below its pKa. Measurements of thermodynamic parameters of proteins in the presence and absence of trimethylamine N-oxide. *J Biol Chem* 280:11035–11042.

55. Miyawaki O (2009) Thermodynamic analysis of protein unfolding in aqueous solutions as a multisite reaction of protein with water and solute molecules. *Biophys Chem* 144:46–52.
56. Holthauzen LMF, Auton M, Sinev M, Rösigen J (2011) Protein stability in the presence of cosolutes. *Meth Enzymol* 492:61–125.
57. Buchner GS, Shih N, Reece AE, Niebling S, Kubelka J (2012) Unusual cold denaturation of a small protein domain. *Biochemistry* 51:6496–6498.
58. Redfield C (2004) NMR studies of partially folded molten-globule states. *Methods Mol Biol* 278:233–254.
59. Pace CN (1975) The stability of globular proteins. *CRC Crit Rev Biochem* 3:1–43.
60. Moon AF, Mueller GA, Zhong X, Pedersen LC (2010) A synergistic approach to protein crystallization: combination of a fixed-arm carrier with surface entropy reduction. *Protein Sci* 19:901–913.
61. Privalov PL (1989) Thermodynamic problems of protein structure. *Annu Rev Biophys Biophys Chem* 18:47–69.
62. Sudol M (1996) Structure and function of the WW domain. *Prog Biophys Mol Biol* 65:113–132.
63. McKnight CJ, Matsudaira PT, Kim PS (1997) NMR structure of the 35-residue villin headpiece subdomain. *Nat Struct Biol* 4:180–184.
64. Neidigh JW, Fesinmeyer RM, Andersen NH (2002) Designing a 20-residue protein. *Nat Struct Biol* 9: 425–430.
65. Scian M, Lin JC, Le Trong I, Makhatazde GI, Stenkamp RE, Andersen NH (2012) Crystal and NMR structures of a Trp-cage mini-protein benchmark for computational fold prediction. *Proc Natl Acad Sci USA* 109:12521–12525.
66. Dill KA, Alonso DOV, Hutchinson K (1989) Thermal stabilities of globular proteins. *Biochemistry* 28:5439–5449.
67. Spassov VZ, Karshikoff AD, Ladenstein R (1995) The optimization of protein-solvent interactions: thermostability and the role of hydrophobic and electrostatic interactions. *Protein Sci* 4:1516–1527.
68. Tsai J, Gerstein M, Levitt M (1997) Simulating the minimum core for hydrophobic collapse in globular proteins. *Protein Sci* 6:2606–2616.
69. Chiu TK, Kubelka J, Herbst-Irmer R, Eaton WA, Hofrichter J, Davies DR (2005) High-resolution x-ray crystal structures of the villin headpiece subdomain, an ultrafast folding protein. *Proc Natl Acad Sci USA* 102:7517–7522.
70. Song J, Park JK, Lee J-J, Choi Y-S, Ryu K-S, Kim J-H, Kim E, Lee K-J, Jeon YH, Kim EE (2009) Structure and interaction of ubiquitin-associated domain of human Fas-associated factor 1. *Protein Sci* 18:2265–2276.
71. Becktel WJ, Schellman JA (1987) Protein stability curves. *Biopolymers* 26:1859–1877.
72. Minton AP (2000) Effect of a concentrated “inert” macromolecular cosolute on the stability of a globular protein with respect to denaturation by heat and by chaotropes: a statistical-thermodynamic model. *Biophys J* 78:101–109.
73. Wang Y, Sarkar M, Smith AE, Krois AS, Pielak GJ (2012) Macromolecular crowding and protein stability. *J Am Chem Soc* 134:16614–16618.
74. Cooper A (1984) Protein fluctuations and the thermodynamic uncertainty principle. *Prog Biophys Mol Biol* 44:181–214.
75. Stams T, Niranjana Kumari S, Fierke CA, Christianson DW (1998) Ribonuclease P protein structure: evolutionary origins in the translational apparatus. *Science* 280:752–755.
76. Henkels CH, Kurz JC, Fierke CA, Oas TG (2001) Linked folding and anion binding of the *Bacillus subtilis* ribonuclease P protein. *Biochemistry* 40: 2777–2789.
77. Chang Y-C, Oas TG (2010) Osmolyte-induced folding of an intrinsically disordered protein: folding mechanism in the absence of ligand. *Biochemistry* 49:5086–5096.
78. Love JJ, Li X, Chung J, Dyson HJ, Wright PE (2004) The LEF-1 high-mobility group domain undergoes a disorder-to-order transition upon formation of a complex with cognate DNA. *Biochemistry* 43:8725–8734.
79. Kuloğlu ES, McCaslin DR, Markley JL, Volkman BF (2002) Structural rearrangement of human lymphotactin, a C chemokine, under physiological solution conditions. *J Biol Chem* 277:17863–17870.
80. Marsh JA, Dancheck B, Ragusa MJ, Allaire M, Forman-Kay JD, Peti W (2010) Structural diversity in free and bound States of intrinsically disordered protein phosphatase 1 regulators. *Structure* 18:1094–1103.
81. Marley J, Lu M, Bracken C (2001) A method for efficient isotopic labeling of recombinant proteins. *J Biomol NMR* 20:71–75.
82. Pace CN, Vajdos F, Fee L, Grimsley G, Gray T (1995) How to measure and predict the molar absorption coefficient of a protein. *Protein Sci* 4:2411–2423.
83. Rees JS, Lilley KS (2011) Method for suppressing non-specific protein interactions observed with affinity resins. *Methods* 54:407–412.
84. Cleveland WS, Devlin SJ (1988) Locally weighted regression – an approach to regression-analysis by local fitting. *J Am Stat Assoc* 83:596–610.
85. Wang A, Bolen DW (1997) A naturally occurring protective system in urea-rich cells: mechanism of osmolyte protection of proteins against urea denaturation. *Biochemistry* 36:9101–9108.
86. Shaw KL, Scholtz JM, Pace CN, Grimsley GR (2009) Determining the conformational stability of a protein using urea denaturation curves. *Methods Mol Biol* 490:41–55.
87. Street TO, Courtemanche N, Barrick D (2008) Protein folding and stability using denaturants. *Methods Cell Biol* 84:295–325.
88. Johnson ML, Faunt LM (1992) Parameter estimation by least-squares methods. *Meth Enzymol* 210:1–37.
89. Kingston RL, Baker HM, Baker EN (1994) Search designs for protein crystallization based on orthogonal arrays. *Acta Crystallogr D* 50:429–440.
90. Otwinowski Z, Minor W (1997) Processing of X-ray diffraction data collected in oscillation mode. *Meth Enzymol* 276:307–326.
91. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) Phaser crystallographic software. *J Appl Cryst* 40:658–674.
92. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D* 66:486–501.
93. Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D* 53:240–255.
94. Pringle CR, The order Mononegavirales: evolutionary relationships and mechanisms of variation. In: Gibbs A, Calisher CH, García-Arenal F, Eds. (1995) *Molecular basis of viral evolution*. Cambridge University Press, Cambridge, pp 426–437.

95. McCarthy AJ, Goodman SJ (2010) Reassessing conflicting evolutionary histories of the Paramyxoviridae and the origins of respiroviruses with Bayesian multigene phylogenies. *Infect Genet Evol* 10: 97–107.
96. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7: 539.
97. Ronquist F, van der Mark P, Huelsenbeck JP, Bayesian phylogenetic analysis using MrBayes. In: Lemey P, Salemi M, Vandamme A-M, Eds. (2009) *The phylogenetic handbook*. Cambridge University Press, Cambridge, pp 210–266.
98. Whelan SPJ, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 18:691–699.
99. Rother K, Preissner R, Goede A, Frömmel C (2003) Inhomogeneous molecular density: reference packing densities and distribution of cavities within proteins. *Bioinformatics* 19:2112–2121.
100. Olechnovič K, Margelevičius M, Venclovas C (2011) Voroprot: an interactive tool for the analysis and visualization of complex geometric features of protein structure. *Bioinformatics* 27:723–724.
101. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera – a visualization system for exploratory research and analysis. *J Comput Chem* 25:1605–1612.