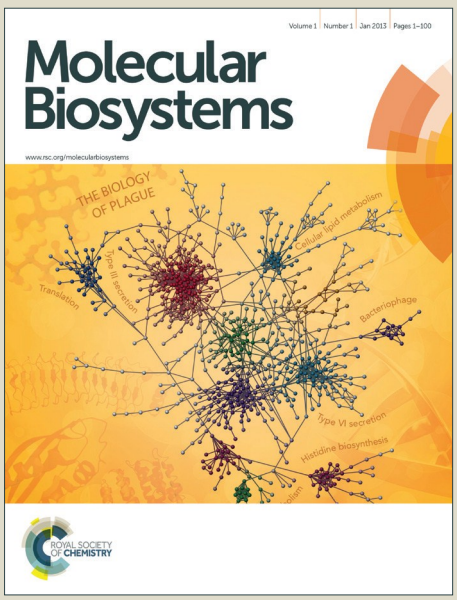
Molecular BioSystems

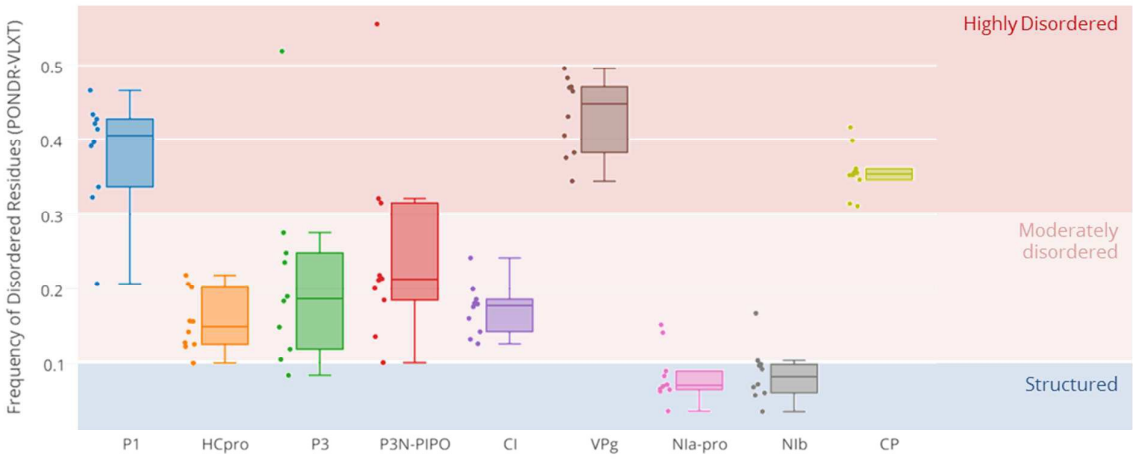
Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

First proteome-wide analysis of intrinsic disorder in a major plant virus genus.



Protein intrinsic disorder within the *Potyvirus* genus: from proteome-wide analysis to functional annotation

Justine Charon[a,b*], Sébastien Theil[a,b], Valérie Nicaise[a,b] and Thierry Michon[a,b*]

[a]UMR Biologie du Fruit et Pathologie, INRA, Villenave d'Ornon cedex, France

[b]UMR Biologie du Fruit et Pathologie, Université de Bordeaux, Villenave d'Ornon cedex, France

Running title : Functions of protein intrinsic disorder in *Potyvirus* genus

*Corresponding authors:

Justine Charon, justine.charon@bordeaux.inra.fr

Thierry Michon, thierry.michon@bordeaux.inra.fr

Abstract

Within proteins, intrinsically disordered regions (IDRs) are devoid of stable secondary and tertiary structures under physiological conditions and rather exist as dynamic ensembles of inter-converting conformers. Although ubiquitous in all domains of life, intrinsic disorder content is highly variable in virus genomes. Over the years, functional annotations of disordered regions at the scale of the whole proteome have been conducted for several animal viruses. But to date, similar studies applied to plant viruses are still missing. Based on disorder prediction tools combined with annotation programs and evolutionary studies, we analyzed the intrinsic disorder content in *Potyvirus*, using a 10-species dataset representative of this genus diversity. In this paper, we revealed that: i) potyvirus proteome displays a high disorder content, ii) disorder is conserved during potyvirus evolution, suggesting a functional advantage of IDRs, iii) IDRs evolve faster than ordered regions and, iv) IDRs may be associated to major biological functions required for the potyvirus cycle. Notably, the proteins P1, Coat protein (CP) and Viral genome-linked protein (VPg) display a high content of conserved disorder, enriched in specific motifs mimicking eukaryotic functional modules and suggesting strategies of host machinery hijacking. In these three proteins, IDRs are particularly conserved despite their high amino acid polymorphism, indicating a link to adaptive processes. Through this comprehensive study, we further investigate the biological relevance of intrinsic disorder in potyvirus biology and we propose a functional annotation of potyviral proteome IDRs.

Importance

Two main biological advantages can be associated to intrinsic disordered regions (IDRs) in virus proteomes. First, IDRs confer a structural flexibility required for the viral proteins to interact with several different partners and ensure their multifunctionality during infection. Second, the low

topological requirements associated to intrinsic disorder allow mutational permissiveness, enabling virus adaptation. In this context, assessing the content and the distribution of intrinsic disorder in a whole viral genus can help to better understand virus-host interactome, and to shed light on adaptive processes leading to host defense overcoming/escape and new diseases emergence. Owing to the importance of protein intrinsic disorder in animal virus biology, we investigated for the first time the occurrence of intrinsic disorder in *Potyvirus,* a major plant virus genus, in an attempt to identify biologically-relevant IDRs and propose functional annotations of these regions as part of potyvirus biology exploration.

Introduction

About 25 years ago, the emergence of the intrinsic disorder concept, which defines proteins or proteins regions as devoid of stable and fixed secondary and/or tertiary structure in physiological conditions and in absence of binding partners, started to challenge the classical view of proteins structure/function relationship [1–8]. Contrary to stable and unique structures found in ordered or globular proteins, intrinsically disordered proteins (IDPs) or intrinsically disordered regions (IDRs) share the peculiarity to exist as dynamic ensemble of conformers in the cell [2–4,9,10]. The key idea behind this discovery, which at the start puzzled many structuralists, was that protein intrinsic disorder had a functional relevance. In this sense, a growing number of studies report the role of disorder in multi-interactions and post-translational regulation processes [11–14], signal transduction and control pathways, as well as involvements in evolutive processes [15–21].

This strong interest of the scientific community encouraged the development of robust *in silico* predictors to estimate disorder content from amino acid (aa) sequences. This contributed over the last years to a high-throughput proteome analysis that established the ubiquitous nature of intrinsic disorder in all domains of life [22,23]. Useful databases are now available, such as DisProt

that lists today more than 700 IDPs and 1500 IDRs for which disorder prediction matches with experimental data [24].

With between 7,3% to 75% of residues being disordered, the proteome of viruses in its whole presents the highest variability of disorder in the living world [22,23,25]. Viruses are proven masters in genome compactness, as most viral proteomes usually encode for a few proteins that must ensure a diversity of functions through multiple interactions. Recently, intrinsic disorder has been experimentally related to these multi-interaction and binding promiscuity in several studies [26–29].

Another aspect of protein disorder in virus lifestyle concerns the extraordinary abilities of these microbes to quickly adapt to various environments. Low topological constraints exerted on disordered regions of some viral proteins have already been experimentally associated to a larger mutational tolerance [30]. This could explain in part how RNA virus can cope with their high mutation rates (10^{-3} to 10^{-5} error/nt/replication cycle)[31]. This mutational robustness is even more surprising given the commonly-used frameshift strategy in viral genomes and the multifunctionality of the proteins they encode. Consequently to this observation, it has been postulated that mutations could accumulate in IDRs without strong deleterious effect on protein stability and function, while facilitating viral adaptation [32–36] and the emergence of new functions [37,38].

In this context, assessing the disorder content and distribution in viral species/families can help to better understand some key-features of virus biology, such as adaptive processes leading to host resistance breaking, defense escape and new diseases emergence. Overall, while bearing in mind the necessary cautiousness inherent to predictive approaches, addressing “disordome” and its functional implications in virus biology can result in finding common traits of disorder

functionality. These findings could in turn feed prediction/annotation programs to better apprehend newly discovered viral families.

Functional annotations of disordered regions at the scale of the whole proteome have already been conducted for some well-studied animal viruses, like *Hepatitis C virus* (HCV) [39], *Human immunodeficiency virus-1* (HIV) [40], *Human papillomavirus* (HPV) [41] and more recently *Dengue virus* [42], with other members of *Flavivirus* genus (*e.g. Yellow fever virus*, *Japanese encephalitis virus* and *West nile virus*)[43].

To date, same kind of analyzes applied to plant viruses are still missing. Nevertheless, strong experimental evidences for the presence of functional disorder in a reduced number of potyvirus proteins have been recently reported [29,44–48].

Potyvirus genus represents one of the largest and the most economically destructive genus of plant viruses. Because of their wide host range and spread/dispersion, potyviruses are very hard to contain and manage, and cause dramatic losses in cultural crops worldwide [49,50]. Among potyviruses, both *Potato virus Y* (PVY) and *Plum pox virus* (PPV) rank in the "Top 10" plant viruses of major social and economic impact [51].

Potyvirus virions are flexuous filamentous particles that include a 10kb positive-sense single-stranded RNA molecule ((+)ssRNA) [52–54]. The viral protein genome-linked (VPg) is covalently attached at its 5’ end and a poly(A)-tail terminates its 3’ end [55–57]. This genome contains untranslated or non-coding regions (UTR) at each of its ends, that surround a single-ORF encoding a polyprotein ~3'000 residues long. After translation, the polyprotein is proteolytically processed by three potyvirus-encoded proteases P1, Helper component proteinase (HC-Pro) and Nuclear inclusion proteinase (NIa-Pro) into ten mature proteins [58–62]. In addition to the ten proteins ensuing from the polyprotein maturation, an eleventh protein, P3N-PIPO, is translated from a +2

ribosomal frameshift in the P3 sequence [63]. The replication step occurs within viral factories localized into intracellular specific membranes, where the nuclear inclusion b (NIb), the potyviral replicase, performs the synthesis of new genomic RNAs [64]. After the translation and replication processes, viral genomes can be addressed to different fates, comprising encapsidation into new particles, degradation, or cell-to-cell, long distance and vector-mediated movement. Advances in the understanding of processes underlying regulation of those pathways are reviewed in [65].

In this context, we propose here to examine the occurrence of intrinsic disorder in the entire *Potyvirus* genus and to assess its involvement in viral functions. Owing to the importance of protein intrinsic disorder in viral functions, performing this analysis on a whole plant virus genus is worth doing. In this work, we assessed for the first time the disordome of *Potyvirus*, through *in silico* characterization of intrinsic disorder at the whole genus level. In addition, an attempt was made to identify biologically-relevant IDRs and establish a functional annotation of these regions as part of the study of *Potyvirus* biology.

Materials and Methods

Potyviral sequences dataset

Potyviral full length genomic sequences of the 10 potyvirus species considered were retrieved from the National Center for Biotechnology Information (NCBI) resource (http://www.ncbi.nlm.nih.gov/).The 5' and 3' non-coding regions (UTRs) were discarded and the remaining genomic sequences were translated into polyproteins using MEGA6.0 software [66]. Sequences containing ambiguous characters were curated. Due to dN/dS calculation requirements, datasets of more than 40 sequences per species were curated to remove redundant or almost identical sequences for preserving the database diversity. A total of 288 non redundant polyproteins was thus obtained (listed in Table S1).

To extract individual proteins sequences, polyproteins nucleotide sequences were aligned and each open reading frame (ORF) was retrieved from its respective GenBank annotations. ORFs were then translated into protein sequences. A different process was conducted to build the P3N-PIPO dataset, given its intra-species length polymorphism [67]. For each species, the sequences ends were manually adjusted. P3 and P3N-PIPO sequences were not included in the dN/dS study, because ribosomal frameshifts introduce too much bias for this type of evolutionary sequence analysis.

Determination of disorder in potyvirus proteins

Disorder prediction

Disorder was predicted both in polyproteins and in each individual proteins using PONDR-VLXT® predictor [68], available on Disprot database resource (www.disprot.org) [24]. This neuronal learning-based predictor employs algorithms using a set of features based on biological knowledge. It was trained on disorder data derived from either X-ray crystallography or nuclear magnetic resonance. It is particularly suitable for proteins that share ordered and disordered regions at the same time, which is typical of viral proteins. This predictor is also very accurate in the detection of sites involved in molecular recognition, signaling and regulation. There are generally good correlations between PONDR-VLXT® predictions and experimentally probed disorder. It is the case for the potyviral VPg, for which predictions are in very good agreement with the experimental characterization of the disorder properties of VPgs originating from several potyviral species [29,45,46]. PONDR-VLXT® predictions are also in agreement with the experimental disorder characterization of the *Potato virus A (PVA)* CP [48]. The software uses a sliding window (set to 7 amino acids length in this study) to associate a disorder score to each residue. Residues with a score below 0.5 were considered as ordered and residues with a score higher than 0.5 are considered as disordered.

Mean disorder content and conservation calculation

For each viral sequence within a given species, the disorder/order state was determined at each residue position. After intra-species sequence alignment, this disorder/order state was compiled. An average score at each amino acid position along viral sequences representative of all sequences was obtained and expressed as a percentage of disorder state conservation at each position. Heatmaps of these conservation scores were built for each individual protein within the 10 species datasets using plotly software (https://plot.ly/).

Segmenting the protein sequences into ordered and disordered regions

Comparison of the evolutionary rates associated to ordered and disordered state was initiated by separating ordered and disordered regions. Within each of the 10 species datasets, PONDR-VLXT® disorder scores obtained for all sequences at each amino acid position were used to calculate average scores resulting in a consensus disorder prediction profile along the polyprotein. On the 10 consensus profiles, ordered and disordered regions were defined as a minimum of 5 consecutive amino acids displaying disorder scores >0.5 and <0.5 respectively. The resulting segmentation patterns were used to cut all the polyprotein sequences within each species datasets.

dN/dS ratio calculations

To evaluate the evolutive constraints exerted on each IDR and ordered regions, we determined dN/dS ratios, defined as non-synonymous nucleotide substitutions (resulting in amino acid changes) vs synonymous nucleotide substitutions (no amino acid changes). To this end, only domains encoded by nucleotide~~s~~ sequences of at least 60 nucleotides long were retained to maintain the analysis robustness [69]. PARtitioning approach for Robust Inference of Selection (PARRIS) method [70] was used (default parameters). It is part of the Hyphy package [71] which is

available on www.datamonkey.org website [72,73]. This method is particularly suitable for viral sequences analysis, which often results from recombination events [74] that could lead to misinterpretation of phylogenetic results [75]. Finally, ω value (corresponding to mean dN/dS ratio) is inferred from all sequences. Mean dN/dS are classified into ordered and disordered domains. A Kruskal-Wallis non-parametric statistical test was applied to mean dN/dS values obtained from IDRs and ordered regions from to detect significant difference of dN/dS -values between those 2 classes.

Phylogenetic analysis

To infer phylogenetic tree of potyvirus, a polyprotein amino acid sequence per species was randomly chosen. They were aligned using the high speed multiple sequence alignment program MAFFT (Multiple Alignment using Fast Fourier Transform) [76] with default parameters. Two outgroups sequences (*Agropyron mosaic virus* and *Hordeum mosaic virus*) belonging to *Rymovirus* genus were also used for multiple alignment, according to [77]. The phylogenetic tree was inferred from polyproteins alignment using Maximum Likelihood method, with MEGA6.0 software. Resulting phylogenetic tree was customized with FigTree software (http://tree.bio.ed.ac.uk/software/figtree/).

Functional annotations of potyvirus proteins

Amino acid diversity score calculation

Amino acid diversity of protein alignment was determined using Shannon entropy measure, available on Protein Variability Server (http://imed.med.ucm.es/PVS/) [78]. It consists in assigning a diversity score H at each alignment position.Typically, position is considered as variable when the H parameter is higher than 2 and conserved when H is below 2. Shannon's scores were displayed

for each species dataset as heatmaps. A blue (low H-score) to yellow (high scores) color gradient was used to visualize diversity along protein sequences, using Plotly program.

Determination of MoRF

Molecular recognition features (MoRF) are predicted from amino acid sequences, using MoRFPred tool [79](http://biomine-ws.ece.ualberta.ca/MoRFpred/), an algorithm fed with a large dataset of annotated MoRF regions. MoRF predictions were done on each protein sequences and aligned. MoRF signals corresponding to less than 5 consecutive residues were considered as not biologically relevant and were discarded. The intra-species MoRF conservation was determined, and represented as a color gradient (green boxes on protein heatmaps Supplemental data S2).

Determination of ELMs

To assess the abundance of motif usage in the virus world, an *in silico* prediction of the occurrence of 173 known functional ELMs was attempted in a dataset of 2,208 non-redundant viruses belonging to various groups and including the 10 potyvirus species used in our study [37]. Given the difficulty to predict motifs with biological relevance, only motifs occurring in IDRs were considered In order to identify potential new functions/interactions associated to intrinsic disorder in potyviral proteins, we extracted ELM predictions obtained for potyviruses from the whole viral ELMs dataset provided by Hagai and co-workers. High probabilities of occurence are inherent to the low complexity of many ELMs. However, given their high biological relevance in the viral context, some of these motifs could not be discarded. Motifs conservation strongly correlates with functionality [80]. Consequently, this criterion was used and we applied a 80% conservation cut-off to discriminate predicted biologically relevant ELM from false-positive ones. Potentially relevant ELMs are listed in supplementary data, Table S2.

224 *Manual annotation of functional domains and proteolytic cleavage sites*

225 Functional annotations of polyproteins and proteins were retrieved for each species both from
226 literature and UniProtKB website (http://www.uniprot.org/uniprot/).

227

228 Results and Discussion

229 *Assessment of disorder abundance in the Potyvirus genus*

230 To assess whether protein intrinsic disorder represents a strategy employed by potyviruses that
231 would be selected during their evolution, we considered 10 potyvirus species representing the
232 diversity of the whole genus [77], and for which more than fifteen full-length genome sequences
233 were available on NCBI GenBank resource (see Material and Methods Section). The chosen species
234 and their respective sequence accessions are given in supplemental data (Table S1). For each
235 species dataset, intrinsic disorder was first predicted at the polyprotein level (Fig. 1) using PONDR-
236 VLXT® as a robust predictor (see Material and Methods Section). The calculated "average number
237 of disordered residues per polyprotein" is close to 20%, whatever the inter-species phylogenetic
238 distance. Importantly, starburst radiation occurring during potyvirus evolution [81] prevented us for
239 going deeper in the study of phylogenetic dynamics of intrinsic disorder in the genus. However,
240 the disorder frequency in potyviruses seems not directly related to the evolutive history of this
241 group, as no disorder enrichment or depletion seems to be correlated to any evolutive lineages
242 (Fig. 1A). The disorder content was homogenous inside each species, as shown by low mean root
243 square deviations. This was also observed at the inter-species level (Fig. 1B). Such an homogeneity
244 is in contrast with the high variability of disorder content observed in some other viral families and
245 despite the long-scale evolutive divergence of the *Potyvirus* genus [25,81]. Interestingly, a higher
246 propensity of intrinsic disorder is observed for *Lettuce mosaic virus* (LMV) species and results from

the particular contribution of the viral protein P3 (discussed in "P3 and P3N-PIPO contain constrained disordered regions" section). Overall, these results indicate that potyvirus belong to the group of viruses displaying high disorder in their proteome, similarly to other (+)ssRNA viruses [22,23,25].

Given the high diversity of this viral genus, the conservation of intrinsic disorder during potyvirus evolution strongly suggests an evolutive hallmark selected during potyvirus evolution and thus a functional advantage of this structural feature.

To assess the intrinsic disorder distribution inside potyvirus proteomes, polyproteins were segmented into their individual protein sequences and PONDR-VLXT® predictor was used to predict the disorder content in the eleven mature potyviral proteins. A box plot representing the disorder variability for nine of the eleven proteins at the inter-species scale is plotted in Fig. 2A, while the variation at intra-species scale is figured in Fig. 2B. Except few exceptions (P1 from PPV and P3/P3N-PIPO from LMV), all potyviral species share the same distribution of intrinsic disorder for each protein (Figs. 2A,B). Based on a previously described method [82], the proteins can be classified into three groups according to their mean disorder propensity (Fig. 2A). NIa-Pro and NIb replicase are classified as "structured" with an overall disorder content of less than 10%. HC-Pro, P3, P3N-PIPO and CI are classified as "moderately disordered" with an intermediate disorder content comprised between 11 and 30%. Finally, VPg, P1 and CP have an overall disorder content globally higher than 30% and are classified as "highly disordered" (Fig. 2A). VPgs from LMV, PVA and PVY were previously experimentally proven as disordered proteins, as well as the CP from PVA [29,44–48]. These data validate the strength of our prediction for these two proteins. On the basis of the disorder analysis in our dataset, it is more than likely that the disorder features of these

viruses can be extended to the other members of the *Potyvirus* genus. This legitimates the hypothesis of a functional assignment of disorder in potyviruses.

Disorder involvement in molecular functions, multi-interactions and adaptation

No direct correlation was identified so far between disorder content and functions in potyviral proteins: for instance, the 5 proteins carrying enzymatic functions (P1, HC-Pro, CI, NIa-Pro, NIb) have variable disorder contents (from P1 highly disordered to NIb highly ordered, Fig. 2A).

It was hypothesized that intrinsic disorder could confer to a single domain the necessary plasticity to fulfill several different functions [83–85]. Nevertheless, using data on potyvirus protein interactome reviewed in [86], we observed that the multi-interactions ability of potyviral proteins is not correlated with their mean disorder content ($R<0.01$; P-value>0.8) (data not shown).

Intrinsic disorder has been related to adaptive abilities of RNA viruses to escape host immunity responses to breakdown host resistances or to enlarge their host specificity [16,30,37]. Interestingly, potyviral VPg, P1 and CP, proteins classified as highly disordered (Fig. 2A), have already been associated to adaptive events (reviewed in [87]). It prompted us to perform a deeper analysis of this adaptation-disorder relationship. However, the examination of published data on some plant virus virulence factors, already defined as directly related to adaptive processes, didn't give any correlation between these factors and their mean disorder content. Additionally, no clear relationship was observed between evolutive constraints and the disorder content in each protein considered as a whole, as reported with their respective calculated dN/dS ratio (or ω) at intra-species level (Fig. S1).

To conclude on this part, no correlation can be made so far between the whole-protein disorder content and general molecular and biological functions. This led us to perform a deeper analysis of intrinsic disorder along the proteins sequences with the aim to identify specific conserved

292 disordered regions and to discuss their possible involvement in known and sequence mapped
293 interactions.

294 ***Functional annotation of IDRs in potyviruses***

295 **Potyvirus proteins IDRs evolve faster than ordered regions.**

296 We reasoned that because of their multiple roles during plant infection, potyvirus proteins could
297 be viewed, in first instance, as tandems of functional modules. Consequently, we conducted a
298 comparison of the evolutive constraint (dN/dS) between ordered and disordered regions along the
299 polyprotein of each virus species. Polyproteins were segmented into ordered and disordered
300 regions according to PONDR-VLXT® predictions. The dN/dS ratio was calculated for all the
301 sequences of at least 20 residues long (see details in Material and Methods section). Globally, in
302 potyviruses, IDRs display significantly higher dN/dS values than ordered ones (Fig. 3), that
303 indicates a true tendency of intrinsically disordered domains to evolve faster than more structured
304 regions during potyvirus evolution. Such weaker evolutive constraints on IDRs strongly suggest
305 abilities for a faster and easier exploration of adaptive solutions and the emergence of new
306 functions in viral proteins [30,37]. Nevertheless, some potyviral IDRs display a low dN/dS (Fig. 3),
307 illustrating the concepts of "constrained" and "flexible" disorder discussed below in section
308 "*amino acid polymorphism"*.

309 **Disorder and proteolytic processing of the potyviral polyprotein**

310 The potyvirus polyprotein is proteolytically processed into ten mature proteins by the three
311 potyvirus-encoded proteases P1, HC-Pro and NIa-Pro [59,61,62](Fig. 4A). The P3N-PIPO protein results
312 from a translation frameshift and cannot be considered as a proteolytic product [63]. P1 and HC-Pro
313 are self-cleaved at their C-terminal (C-ter) end and NIa-Pro cleaves at 7 different sites in the
314 polyprotein central and C-ter regions [88] (Fig. 4A). The potyvirus genome strategy to encode for a

single ORF has been related to a way of tuning the proteins function through a sequential proteolytic process during the viral cycle [58]. Indeed, functional relevant intermediary forms have been identified as key players in viral infection and adaptation. The 6K2-VPg-NIa-Pro intermediate is anchored to the ER membrane during potyvirus replication [89–91]. The 6K2 proteolytic separation from CI and VPg modulates viral replication rate and movement [58]. Deleterious effects on viral fitness of NIb relocation at different positions in the polyprotein also highlight the functional relevance of proteins placement within the polyprotein and the importance of transient intermediary forms in the viral cycle [92]. The cleavage efficiencies by NIa-Pro at each of its cleavage sites was investigated. Most proteolytic sites were fast processed (6K1/CI, 6K2/VPg, NIa-Pro/NIb and NIb/CP) but three of them, namely P3/6K1, CI/6K2 and VPg/NIa-Pro, were processed at a lower rate [58]. These data strongly suggest that differences in cleavage efficiencies at various NIa-Pro sites are of strong functional relevance. A fine time-dependent tuning of the polyprotein maturation is likely to be important for the virus biology. For instance, the post-translational phosphorylation of NIa-Pro was demonstrated to inhibit the trans-proteolytic cleavage of VPg-Pro (Mathur & Savithri 2012).

The site accessibility to the enzyme strongly influences the proteolytic cleavage efficiency. It is well-established that intrinsic disordered regions undergo faster proteolytic digestion than more structured ones [94–97]. In viral polyproteins, the propensity to disorder in the vicinity of cleavage sites has already been predicted for *Hepatitis C virus* [39] and *Dengue Virus*. Remarkably, our comparative analysis of the disorder propensity at the inter-species scale shows that the NIb-CP cleavage site is located in a disordered segment in all species (Fig. 4B). The conservation score of the intrinsic disorder state along the polyprotein was also determined for the ten potyvirus species dataset (Fig. 4B). Disorder was considered as conserved intra-species when it was observed in at least 80% of the sequences (Fig. 4B). P3/6K1, CI/6K2 and VPg/NIa-Pro proteolytic sites, which were

demonstrated as slowly processed [58] share a low score of disorder conservation between 0% and 40% (Fig. 4A). By contrast, disorder is conserved in some of the other NIa-Pro cleavage sites with up to 100% conservation in 6K2/VPg and NIb/CP sites (Fig. 4B). The high release rate observed for CP [58] is directly correlated with a high conservation of disorder at this cleavage site and well-supports the hypothesis of protease-sensitivity of unstructured/flexible regions. The genome encapsidation requires a large pool of monomeric CP units, and it is likely that a fast release of CPs facilitates the assembly. In addition, it was recently proposed that the pool of free CP could finely tune the steps of genome translation and encapsidation [98]. This seems to illustrate the evolutive constraint exerted on disorder status because of its role as proteolytic "facilitator". CI-6K2, is a stable intermediate form *in vivo,* and has been associated to replicative functions in the viral cycle. It is noteworthy that the cleavage site is part of a conserved ordered region in 100% of the potyviruses considered here (Fig. 4A,B). The prevalence of a structural order in this proteolytic region could slow down the proteolysis by NIa-Pro, resulting in a longer availability of this intermediate required for the viral cycle. Conversely, despite of a relatively low disorder-state conservation of NIa-Pro/NIb sites among potyviruses, the proteolytic release of the corresponding intermediary forms is fast [58]. In this case, intrinsic disorder does not seem to be related to any proteolytic facilitation. To conclude, only partial correlations can be established between order/disorder states of the cleavage sites within the polyprotein and their processing kinetics by NIa. It is likely that the order-disorder balance in the vicinity of the polyprotein maturation sites provides a kinetic modulator participating to the necessary chronologic apparition of various functional intermediates during the potyviral cycle. In this respect, the self-release of P1 from HC-Pro requires a special discussion (see "P1: Polymorphism, but conserved IDRs" section).

361 ***Protein-by-protein analysis of IDRs functions***

362 Although it is difficult to associate predicted conserved IDR to known biological processes and
363 functions, the task deserves to be done, as structural disorder proves to be a main strategy for
364 many viral functions [27,35,39]. To this end, a functional annotation of the potyviral IDRs found within
365 each protein was performed through an analysis of: i) their amino acids polymorphism, ii) their
366 content in recognition-binding motifs and, iii) their content in eukaryotic functional linear motifs.
367 The relationship between these three parameters and IDRS is presented as follows.

368 *Amino acid polymorphism analysis.* In many organisms, IDRs are more tolerant to aa changes than
369 structured regions, a statement that seems to be also the case for potyviruses (Fig. 3). Bellay and
370 colleagues proposed that it allows sorting conserved disordered regions in two classes according
371 to their evolutive rates: “flexible disorder“ for disordered regions which display a relatively high
372 amino acid polymorphism (AAP), and “constrained” disorder which refers to disordered regions
373 that are not variable in their aa sequences [19]. Flexible disorder is, in the first instance found in
374 flexible spacers between functional domains [99]. In addition, flexible disorder is particularly
375 observed in proteins with one or two regions interacting sequentially with several partners
376 through conformational switches, termed "date Hubs" [19,100,101]. Constrained disorder, by contrast,
377 is found in IDRs, whose function is linked to more topological requirements. This is the case of
378 multi-interface hubs (named “party hubs”) [100,102,103], that can be involved in simultaneous binding
379 with various partners. Therefore, an analysis of sequence conservation in potyvirus IDRs is
380 expected to give insights on underlying interaction modes.

381 *MoRFs analysis.* Molecular recognition features (MoRFs) are defined as short motifs (5 to 25
382 residues long) embedded in IDRs that undergo disorder-to-order transition upon binding to a
383 partner, according to the induced-folding process [104,105]. MoRFs can be classified in three classes,

depending of the secondary structure resulting from the folding upon interaction: α-MoRFs fold into α-helixes, β-MoRFs into β-sheets and coil-MoRFs into random coils. Functional MoRFs are frequent in viral IDRs [106]. Sequence characteristics (charge, hydrophobicity, conservation and secondary structure properties) allow the prediction of potential MoRFs and help to their functional annotation [79,107,108]. This analysis was performed on our potyviral protein datasets using MoRFPred predictor [79], to detect highly-conserved MoRF motifs potentially involved in interactions. Their relevance and possible involvement in potyvirus biology is discussed on a "case-by-case" basis in view of already documented interactions and the evolutive conservation they display at intra-species level.

ELMs analysis. Eukaryotic Linear Motifs (ELMs) consist in 3 to 10 residues long motifs, which are ubiquitous in eukaryotic organisms. According to the ELM database resource [109](http://elm.eu.org/), they are involved in many interaction-based processes and divided into six functional classes : ligand-binding sites, post-translational modifications, targeting sites, proteasome degradation targeting sites, docking sites and proteolytic cleavage sites. ELMs are often located within IDRs [111], which, with their flexibility, facilitate ELM-based interaction [13,37,112,113]. Being crucial in the emergence of new functions during eukaryotes evolution, ELMs constitute probably a “Achilles heel” since they have been mimicked by viruses during their evolution [113,114]. The use of such eukaryotic short motifs to hijack the host cellular machinery may explain, at least partially, how viruses establish so many interactions despite their high genome compactness. To date, 46 ELMs functional classes among the 219 of the ELM database have been identified in the proteome of viruses from various families, suggesting that ELMs mimicry is a common strategy in the virus world [113]. To assess the abundance of ELM usage in the virus world, an *in silico* prediction of the occurrence of 173 known types of functional ELM was attempted in 2’208 not redundant viruses of diverse groups, including the 10 potyviruses species used in the

present study [37]. Therefore, we selected the potyviral ELMs located in IDRs from the data produced by Hagai *et al.,* in order to identify potential new functions/interactions associated to intrinsic disorder. Given the low complexity and high degenerated nature of these motifs, false positives are expected. In this respect, the localization of conserved ELMs in IDRs constitutes a biological relevant criterion to discriminate true positive motifs. All motifs predicted in potyviral polyproteins share a high probability of occurring by chance. However, given their high biological relevance, some of these motifs could not be discarded. Thus, we applied a "conservation cut-off" to discriminate biologically relevant ELM from false-positive ones. Because motifs conservation strongly correlates with functionality [80], this criterion was used to identify strong relevant ones. Consequently, all predicted motifs found in at least 80% of the polyprotein IDRs from the 10 species considered are listed in Table S2. Among these motifs, only those which were found relevant regarding plant-virus biology were considered (Bold Green boxes in Table S2) and discussed.

P1 : a both highly disordered and variable protein

All potyviral P1 display long and conserved IDRs. The P1 protein, localized to the N-terminus (N-ter) of the potyviral polyprotein, is a serine endopeptidase. It is self-released by cis cleavage at its C-ter [61]. A trans activity was also reported in *Tobacco etch virus* (TEV) [115]. At the inter-species level, disorder was observed in all P1 (Fig. 2). However, P1 has evolved through recombination and duplication events [116] and displays the highest variability both in length (30-63 kDa) and in aa sequence of all potyviral proteins. This prevents the alignment of P1 sequences at the inter-species level. Therefore, such a comparison of the distribution of disorder along the protein was not possible. At the intra-species level, long and conserved IDRs were identified (Fig. 5), suggesting that the conservation of P1 intrinsic disorder in the course of potyvirus evolution is associated to biological functions.

P1 IDRs could contain RNA binding sites. RNA and DNA binding has often been reported as associated to basic positively charged IDRs [117], which act as binding facilitators [118]. The predicted central long-IDR of *Turnip mosaic virus* (TuMV) contains a broad specific RNA binding domain, spanning from residues 150 to 168, that was experimentally confirmed [119]. It is noteworthy that, although highly diverse in sequence at the inter-species level, the first 180 aa of P1 conserve a high averaged isoelectric point value of 9.2, supporting the presence of a nucleic acid binding function. Hence, IDRs in P1 could favor RNA-binding and participate to viral replication and translation processes. TEV P1 physically interacts with the host 80S cytoplasmic ribosomes in infected cells and stimulates translation of viral proteins *in vitro* [120]. The RNA-binding ability of P1 IDRs may be related to RNA-chaperone activity, for which intrinsic disorder is an hallmark [121]. Moreover, such RNA chaperone activity was already reported in other viral proteins [122].

N-ter IDR is expected to modulate P1 proteolytic activity. P1 N-ter domain modulates its proteolytic self-release from the polyprotein, thereby preventing early host defense responses, that would be detrimental to virus systemic infection [123]. The auto-inhibitory effect of intrinsically disordered domains was already reported [124]. Thus, IDRs within the N-terminal part of P1 could fold back to interact with the distant cleavage site at the C-ter end of the protein, slowing down its release. Post-translational modifications, such as phosphorylation, and/or its interaction with another viral/host protein [125] could act as an activator switch.

MoRF in P1. Some well conserved MoRF were found, embedded in P1 disordered regions. One of them, located in a C-ter IDR was shared by 6 of the 10 potyvirus species considered in this study. However, as there is no functional annotation available for this part of the protein, no putative function could be associated with these motifs.

P1 disorder and adaptation. P1 is the most variable potyviral protein in terms of sequence length and AAP. Its involvement in adaptive process and host range specificity was previously reported [116,126,127] and is strongly suggested by its high mean dN/dS (Fig. S1). Interestingly, most of P1 IDRs are well-conserved at the intra-species level (white to red bars, Fig. 5) despite the relatively high diversity of their aa sequence (blue to yellow gradation, Fig. 5). This strongly suggests that: i) there is an evolutive constraint for the conservation of disorder state at these positions that is likely of functional relevance, and ii) these regions are prone to aa switches occurring during virus evolution. This kind of evolutive behavior in IDRs has been referred to as "flexible disorder" [19], which relates to regions binding different partners sequentially. The high variability of sequence in P1 (considered as the least conserved protein among potyviruses), associated to its high degree of disorder strongly support the hypothesis of mutational robustness and lower-evolutionary constraints effect related to disorder, a feature already observed in various proteins [128,129]. This could be related to adaptive abilities to various environments [16], a disorder-based feature already discussed for RNA-viruses [30]. In this context, P1 has already been related to host range specificity within the *Potyviridae* family [116,126,127], reinforcing the idea that disorder could act as an enhancer of adaptation.

P1 disordered regions contain ELMs.

Phosphorylation motifs. Post-translational motifs such as phosphorylation sites are abundant in P1. In disordered regions, surface accessible serine and threonine residues are more prone to phosphorylation than in ordered regions [11–14]. In many cases, these phosphorylations have been reported to modulate functions by preventing or conversely potentiate the folding of intrinsically disordered regions [11]. According to DISPHOS (Disorder-Enhanced Phosphorylation Sites Predictor),

476 at least four serines and two threonines are highly susceptible to phosphorylation, all located in P1
477 N-ter IDR.

478 *14-3-3, FHA and WW binding motifs.* These well-conserved eukaryotic motifs are phospho-sensors
479 consisting in 6-8 residues with a conserved phospho-serine or threonine binding site. 14-3-3
480 proteins (also called GRF for *General Regulatory Factors*) are involved in many regulatory
481 processes in plants (*e.g.* metabolism, hormone signaling, cell-division, stress responses) through
482 their interaction with more than 300 potential targets [130,131]. Several 14-3-3 binding motifs are
483 present within P1 IDRs. The hijacking of that kind of motifs by potyviruses could have many
484 implications in the viral cycle. It was observed that plants deficient for GRF6 display enhanced
485 resistance to PPV infection [132]. Furthermore, these authors reported that GRF6 degradation by the
486 proteasome is stimulated upon virus infection and contribute to plant defense mechanisms. With
487 14-3-3, forkhead-associated (FHA) domains are phosphobinding domains involved in numerous
488 signaling processing, such as metabolism and plant development [131]. FHA-binding motifs have
489 already been characterized as functional motifs in plants [133]. They are conserved in potyvirus P1 as
490 disorder-embedded ELM, and co-localize with predicted phosphorylation motifs (Table S2). The
491 WW domains are known to recruit regulatory protein complexes in various signaling networks.
492 They bind to WW binding motifs, which are short proline or phosphoserine- phosphothreonine-
493 containing motifs. WW-domain proteins have recently been reported as inhibitors of the
494 replication of a (+)ssRNA virus, the *Tomato bushy stunt virus* (TBSV, *Tombusviridae*) [134]. WW
495 binding motifs were predicted in P1 of all species studied here (Table S2). This could provide
496 another example of disorder-mediated ELM mimicry by potyviruses to recruit host factors. All
497 together, these data suggest that FHA, 14-3-3 and WW binding motifs, directly associated to
498 posttranslational phosphorylation sites in P1, are involved in the viral cycle regulation and/or host
499 defense counteracting.

USP7 binding motif. The Ubiquitin Specific Protease USP7 is a member of the large DUB family (deubiquitinating enzymes) in Eukaryotes. It catalyzes the ubiquitin removal from proteins, preventing their degradation by the proteasome. Several examples have been reported in plant virus [135]. Viral DUB activities are encoded by the (+)ssRNA virus *Turnip yellow mosaic virus* (TYMV, *Tymoviridae*) and are involved in controlling levels of RNA-dependent RNA polymerase (RdRp) and viral infectivity [136]. It is likely that controlling the host ubiquitination pathway is a prerequisite for the virus cycle. There is no report of ubiquitin protease activity encoded by the potyvirus genome but the presence of conserved USP7 binding motifs in P1 argues in favor of a possible recruitment of a deubiquitinating activity.

NLS motif. Nuclear localization signals (NLS) motifs are predicted in P1 IDRs. The propensity of disordered regions to display NLS has already been reported [137]. An active nucleolar localization signal (NoLS) was found in the P1 of TEV within a region encompassing residues 50 to 115, which is predicted as disordered in most of the potyvirus species considered in our study. This NoLS seems to be functional, as during the early stage of infection, P1 is found in the nucleolus, where the pre-ribosomal particles processing takes place. Moreover, the protein has been demonstrated as trafficking between the cytoplasm and nucleolus during infection, and to interact physically with the cytoplasmic 80S ribosomal subunit, arguing in favor of its involvement in viral translation [138]. Additional nuclear localization signals are predicted for the majority of potyvirus species. Interestingly, among these conserved ELMs (Table S2), some are not strictly localized in the same IDR. This could suggest a *de novo* apparition of these motifs during potyvirus evolution. Such an evolutionary convergence illustrates both the strong functional benefit underlying these motifs and the evolvability of the corresponding IDRs [13,37,113].

In vitro experimental approaches like disorder characterization through limited proteolysis and secondary structure analysis through circular dichroism should be now undertaken to further investigate the disorder features of this protein. The biological relevance of the detected ELMs could also be experimentally assessed *in vivo* (for instance by site-directed mutagenesis).

HC-Pro, a weakly-disordered multifunctional protein

Next to P1 on the polyprotein, HC-Pro is a cysteine endopeptidase of approximately 460 aa long, which self-cleaves at its C-terminus. HC-Pro is a good example of what virus evolution can design in terms of genome economy. This protein is involved in many processes, such as genome replication, aphid transmission, virus-induced gene silencing, virus polyprotein maturation and virus migration within the plant [139]. According to the electron microscopy of HC-Pro two-dimensional crystals, the N-ter and C-ter domains of the protein are separated by a flexible constriction [140]. Three structural domains can be distinguished: the N- and C-ter regions (approximately 100 aa long each) and the central domain (approximately 250 aa long). The C-ter domain is responsible for the proteolytic activity of the protein. The N-terminal domain is required for aphid transmission of the virus, but most of the HC-Pro functions are located in the central region of the protein. As a whole, only relatively short disordered intra-species conserved regions are predicted in HC-Pro (Fig. S2A). These regions are spread all along the protein sequence and are not conserved between species with the exception of region 161-182 which is predicted as disordered in the 10 species (Fig. S2A). Although resistant to trypsinolysis, HC-Pro from LMV displays two highly exposed sites after Thr170 and Gly176 [140], a feature consistent with this predicted (161-182) IDR (Fig. S2A). This region contains a highly conserved FRNK box allowing the binding of siRNA and miRNA duplexes [141] and has been described as involved in the binding by HC-Pro from *Zucchini yellow mosaic virus* (ZYMV) of the small RNAs methyltransferase HEN1 [142]. Taken together, these features are likely to be related to the gene silencing suppression activity of HC-

Pro in the infected cell. This is another example of the potential role of intrinsic disorder in RNA binding processes. This ability of HC-Pro to interact with elements of the silencing pathway well-illustrates a defense and counter-defense interplay between potyviruses and their hosts. Finally, in most of the species considered, a very short segment (at most 10 residues long) is predicted as disordered within the protease domain (Fig. S2A). In the case of TuMV, the 3D structure of this protease domain, which span residues 301-458, was recently solved [143]. It is mostly structured with the exception of a loop (residues 419–426) that was roughly modeled due to weak electron density. However, for TuMV, this loop is not predicted as disordered. This confirms that many short flexible loops on proteins surface should be differentiated from truly disordered segments.

Disorder in P3 and P3N-PIPO: a mediator of proteins scaffolding at the membrane surface

In spite of its proven involvement in pathogenicity and symptomatology [87,144], P3 remains one of the less characterized potyviral proteins. It has been associated to cell-to-cell movement and to the formation of replication complexes, through its interaction with the ER and replication vesicles [145,146]. P3 of many potyvirus species contain a N-ter conserved hydrophobic region potentially involved in membrane interactions [146]. This region (between the residues 40 and 80) is preceded by an IDR of variable length, which is conserved intra-species. This IDR-containing N-ter part, is important for addressing the protein to the Golgi apparatus [145]. In membranous proteins, an intrinsically disordered tail is commonly observed, protruding in the cytoplasm that is involved in the addressing of various factors at the membrane surface [147]. This P3 IDR could participate in the anchorage of replication vesicles on the ER or alternatively, in the virus addressing to plasmodesmata through interactions with the cytoskeleton. Indeed, P3 traffics along actin filaments and colocalizes with replication vesicles [145,146]. It is noteworthy that the region including this IDR and the adjacent hydrophobic domain is a common feature of P3 and P3N-PIPO, as the

frameshift generating P3N-PIPO is positioned downstream. It could be that both proteins share some functional specificity.

On a broader level, the disorder analysis in P3 showed that there is no clear reproducible IDR pattern at the inter-species level. Nonetheless, some IDRs are highly-conserved intra-species and associated to a low AAP (Fig. S2B). This enables to qualify P3 disorder as globally constrained. Hence, the protein displays features of "party hubs", namely scaffold proteins which interact simultaneously with several partners and/or anchor them to membranes. It is worth mentioning that in LMV, more than 30% of the P3 sequence is predicted as disordered with a succession of large and well-conserved IDRs interspaced by shorter ones. A long N-ter IDR (more than 80 residues) is associated to a highly conserved MoRF, a potential interacting domain with other factors. Interestingly, for LMV, the N-ter hydrophobic region was reported around residue 60 [146]. According to our data, it is embedded in this long IDR (Fig. S2B). This very local hydrophobic signal is typical of MoRF signatures [79] although undetected by MoRFPred.

A 30 aa-long conserved IDR is observed around residue 100 in *Soybean mosaic virus* (SMV) and TuMV (Fig. S2B) that matches with a RubisCO-interacting domain [148]. Finally, a central and relatively conserved IDR is identified between residues 200 and 250 of *Bean Yellow mosaic virus* (BYMV), *Sugarcane mosaic virus* (SCMV), LMV, PPV, PVY and TuMV. This region cannot be functionally annotated because of the absence of detectable ELMs.

P3N-PIPO was recently discovered and results from a +2 ribosomal frameshift in P3 [63]. Its length is highly variable among potyvirus species and has recently been associated to host-driven specificities [67]. In the infected cells, P3N-PIPO localizes in plasmodesmata, suggesting its involvement in virus cell-to-cell movement [149,150]. The PIPO domain interacts with PCaP1, a cation-binding protein anchored to membranes through myristoylation [151]. A putative disordered-embedded MoRF is found in the C-ter region. In spite of its heterogeneity in length, and as

opposed to P3, the IDRs distribution in the PIPO part is conserved throughout in the 10 species. In addition, similarly to P3, IDRs in the PIPO part can be classified as constrained disordered regions. Like P3, P3N-PIPO could act as a party-hub. As a matter of fact, P3N-PIPO interacts with CI [149]. The IDRs within P3N-PIPO could potentiate the simultaneous interactions of the protein with CI and PCaP1. Although overlapping reading frames are frequent and strategic for virus genome economy, they are expected to be heavily impacted by viruses high mutational rates. However, many IDRs better tolerate mutations than structured regions. Interestingly, intrinsic disorder was found to be abundantly generated in viral overlapping reading frames [152]. Disorder both in the PIPO and P3 C-ter part, (orange dashed lines in Fig. S2B,C) observed for most of the studied potyviral species, could prevent destabilizing effects of mutations on these overprinting regions. No ELM was found associated to disorder neither in P3 nor in the PIPO part.

Given their small size (about 50 aa each), and owing to the very short disordered segments predicted for 6K1 and 6K2 and obtained on in their N- and C-terminal parts (data not shown), these two proteins were discarded from our analysis.

CI, a protein with two ordered domains interspaced by a conserved IDR

The cylindrical inclusion protein (CI) is a viral helicase. It interacts with various viral and host partners and is involved in movement and replication (see (147) for review). As HC-Pro, P3 and P3N-PIPO, CI is classified as moderately disordered. This is consistent with CI main functions devoted to catalysis, through its ATPase and RNA-helicase activities, which require well-ordered catalytic domains. The N-terminal region of CI carries the helicase activity [154] and the C-terminal domain possesses binding sites for (at least) two different proteins, the viral VPg, and the plant eIF4E, an initiation translation factor [155]. These two domains are interspaced by a flexible region

(Fig. S2D). This less-structured region is the main contributor to the overall disorder scores obtained (Fig. 2). This IDR is conserved both intra- and inter-species and is roughly located between residue 330 and 370. In the IDR, is included an arginine-rich motif constituting a RNA-binding domain, which corresponds to the most carboxy-terminal conserved domain of the RNA helicases of the superfamily SF2 [156]. In PPV, this RNA-binding domain was mapped in between aa 350 and 402 [154]. Interestingly, such motif is also the hallmark of eIF4A, another SF2 helicase belonging to the translation initiation complex eIF4F highly conserved in all eukaryotes. PONDR-VLXT® prediction obtained for eIF4A displays the same general profile as that of CI, with an ordered background interrupted by short and spaced disordered signals all along the sequence [157]. Hence, it is noteworthy that CI shares with eIF4A a similar intrinsically disordered RNA binding domain, sustaining the hypothesis of a common function associated to disorder.

VPg contains flexible disorder

In potyviruses, VPg constitutes the N-ter part of the NIa protein. After cleavage, the VPg consists in a 22 to 25kDa protein. VPgs are not restricted to phytoviruses, being also involved in the genome replication and protein translation of animal (+)ssRNA viruses. VPg is covalently linked to the viral RNA 5'-end through a conserved tyrosine residue [158,159]. This feature is likely involved in many functional aspects, such as aphid transmission and uncoating process regulation, movement of viral RNA through plasmodesmata, replication initiation through urydilyl VPg priming or the initial steps of translation [90,91,160–171]. VPg can be considered as a hub protein interacting with both various host and viral factors, as well as with the viral RNA [172]. The best documented of these interactions concerns the eukaryotic translation initiation factors eIF4E and eIF(iso)4E [91,173–177]. This association is associated to various tasks during potyvirus infection such as translation, replication and cell-to-cell movement [168,178,179]. Interaction with nucleolar Fibrillarin has also been reported [180], suggesting a function in RNA-silencing or host gene expression regulation. Interaction

641 of VPg with a host RNA helicase RH8, a poly(A)-binding protein (PABP) and the eukaryotic
642 elongation factor 1A (eF1A) clearly demonstrate the involvement of VPg in viral RNA replication
643 and translation [91,176,181,182]. Unfortunately, these interactions have not yet been mapped on the
644 VPg sequence. Intrinsic disorder in VPgs was clearly demonstrated by experimental results
645 obtained in several plant viruses comprising potyviruses and sobemoviruses [29,44–47].

646 IDRs were predicted in all the potyviral VPgs of our dataset. This intrinsic disorder is highly
647 conserved in intra- and inter-species. Conserved IDRs were found at the N-ter and C-ter of
648 potyviral VPgs. The N-ter IDR interacts with PVIP [169], this interaction being linked to virus
649 movement. Downstream, a short highly conserved IDR, centered around residue 50, contains
650 functional NTP-binding and RNA-binding sites [47]. The central part of the VPg displays a conserved
651 IDR about 20 to 30 residues long. This region is involved in the interaction of VPg with eIF4E and
652 HC-Pro [183]. Interestingly, although no MoRF was predicted in this region, the IDR spanning residues
653 89-105 folds into an helix upon binding with eIF4E [44]. At the intra-species level, potyviral VPgs do
654 not share a global high dN/dS value (Fig. S1). But in most of the species, the VPgs central and C-ter
655 IDRs are associated with a high AAP (Fig. S2E). Therefore, the disorder contained in potyviral VPgs
656 is typically of flexible nature. In this respect potyviral VPg is likely to constitute a typical example of
657 “date hub”, with its central IDR interacting sequentially with several partners.

658 Interestingly, the central domain of VPg contains the molecular determinants responsible for the
659 overcoming of many eIF4E-mediated resistances [177,184–187]. As a matter of fact, VPgs shares
660 episodic and localized events of positive selection mostly localized in this central IDR. This leads us
661 to hypothesize about the involvement of disorder status in such adaptive process. The central part
662 of PPV VPg is not predicted as disordered and represents a puzzling exception.

Phosphorylation motifs. As in P1, well-conserved phosphorylation motifs are predicted in VPg disordered regions (Table S2). Both *in vitro* and *in vivo* experiments have already highlighted that VPg is a highly phosphorylated protein [188,189]. However, no functions have yet been associated to these modifications.

Nuclear localization motifs. The N-ter IDR of TEV and PVA VPgs display bi-partite NLS signals [180,190]. Their biological relevance was validated by VPg nucleus localization experiments in several potyviruses [176,180,190]. Functions associated to this nuclear localization still remain unclear but could be related to an hypothetic involvement in host silencing suppression process [180]. Such VPg NLS motifs are retrieved in 8 of the 10 species studied here, reinforcing the idea that VPg nuclear localization is shared among all potyviruses and is very probably associated to important functions in viral cycle.

NIa-Pro displays an inconstant functional C-ter IDR

The 240 residues long C-ter domain of NIa, called NIa-Pro, is a protease, which self-releases from NIa. Its functions in the polyprotein processing have been discussed above (see “Disorder and polyprotein processing” section). Additionally, NIa-Pro possesses a DNase activity possibly involved in regulation of host gene expression [191].

A central IDR between residues 95 and 125 is predicted in half of the species examined. The corresponding region in the TEV and *Tobacco vein mottling virus* (TVMV) proteases is part of a well-resolved surface loop in the 3D structures [192–194]. It cannot be excluded that this region is stabilized in the crystal packing. A short IDR (about 20 residues long) is only predicted at the C-ter of SCMV in our dataset (Fig. S2F). This inconstant IDR was experimentally confirmed in the structures of the TEV and TVMV NIa-Pro. Interestingly, this IDR increases the proteolytic activity in the case of TEV, while it reduces it in the case of TVMV, suggesting this IDR behaves like a

686 modulator of the enzyme [193]. This illustrates how IDRs could tune the viral functions according to
687 the host specificity.

688 **The NIb replicase protein is ordered**

689 NIb is the potyviral replicase, ensuring virus genome amplification. NIb is addressed to membrane-
690 associated viral factories and participates to the formation of replication complexes, through its
691 interaction with VPg, NIa-Pro and several host factors EF1A, Hsc70-3 and PABP [182,195,196]. In spite of
692 the ordered nature of this protein, a well-conserved IDR spanning residues 400 to 450 was
693 predicted in nine of the species. It was not possible to functionally annotate this region. An α-
694 MoRF is predicted for most of the species spanning residues 320-340. However, the low disorder
695 content in the region doesn't support this prediction.

696 As many others viral RdRP, the potyviral NIb displays well-ordered and conserved domains folding
697 into the typical "right hand" structure [197]. However, the low disorder content observed for
698 potyviral RdRP does not strictly constitute a hallmark in RNA virus world. For instance, HCV RdRP
699 displays a mean disorder content of 19,1% [39].

700 **Structural Coat protein (CP), flexible disorder and adaptation**

701 Potyvirus CPs are filamentous particles with helical symmetry made up from the supramolecular
702 assembly of about 2000 CP subunits [54]. According to structural studies, many viral CPs folding
703 consist in a central globular part with N- and C-ter extended arms, found as flexible and disordered
704 at least in the CP monomeric form [198]. These extended arms have been related to several
705 functions, such as nucleic acids interactions, regulation and control of virion assembly and
706 stabilization. In addition to its role in viral genome protection (encapsidation), potyviral CPs also
707 participate to various non-structural functions such as viral RNA translation, replication and
708 movement [199,200].

709 At the inter-species scale, potyviral CPs are heterogeneous in size (from 270 to 350 residues). But
710 all species share the same organization, consisting in: i) a highly conserved ordered central core
711 flanked by a long N-ter IDR, that presents a high intra-species sequence polymorphism, and ii) a
712 more conserved C-ter domain displaying altered short ordered and disordered segments (Fig. S2H)
713 [201,202]. When assembled, the central globular part of the CP, which is inaccessible to proteolysis,
714 forms a structural core interacting with the genomic RNA. The CP N-ter part is accessible from the
715 particle surface as shown by its sensitivity to proteolysis [203]. In PVA, tritium bombardment gave
716 evidence that residues 1-15 and 27-50 are exposed to the surface with at least the 8 first residues
717 being disordered [204]. Depending on the virus species, the 18 to 20 C-ter aa are also exposed to the
718 surface [203]. Importantly, intrinsically disorder was experimentally detected in the CP within the
719 assembled particle [48]. The N-ter region is likely to be the main contributor to this observed
720 intrinsic disorder. The N-terminal part is not structurally essential for the capsid as its deletion has
721 little effect on the virion morphology [205]. This suggests that the structurally flexible N-ter region
722 could participate to non-structural functions in the viral cycle. The N-ter displays a DAG motif, that
723 is involved in the aphid-transmission process [206] by mediating the CP interaction with HC-Pro [207].
724 This interaction site co-localizes with an inter-species highly conserved MoRF (Fig. S2H),
725 supporting plausible induced folding events associated with HC-Pro binding. This N-ter IDR could
726 behave as an extended arm, that explores a large area (according to the “fly-casting” model) and
727 folds as it approaches the actual binding site [208,209]. Contrarily to the N-ter part, which is always
728 exposed at the particle surface, the C-ter region surface exposure varies between potyviral species
729 [203,210]. However, in all species, disorder is predicted in this part (Fig. S2H). A conserved MoRF signal
730 is associated to these IDRs but no interaction was yet reported in this region.

731 *CP, amino acid variability and potyvirus adaptation.* In the N-ter region of potyviral CPs, the high
732 content in conserved disorder is associated with a high amino-acid polymorphism (Fig. S2H). This

illustrates the relationship between low structural constraints and high variability in disordered regions. Given its N-ter extreme aa diversity (revealed by its high mean dN/dS value on Fig. S1), CP is expecting to be a determinant of potyvirus adaptation. Indeed, host-specific determinant motifs in the CP N-ter region have been reported in *Watermelon mosaic virus* (WMV) [211].

ELMs in CP. PVA CP is subjected to phosphorylation. This post-translational modification is directly involved in the RNA-binding regulation, as it reduces the CPs affinity for nucleic acids [212,213]. In PPV, CP phosphorylation and O-GlcNAcylation modifications have also been reported in the N-ter of the protein [214,215]. The CPs phosphorylation state seems to have an enhancing impact on viral infection [213]. These phosphorylation sites were detected in most of the species of our dataset (Table S2). A fine tuning of the CPs pool is required within cells [199]. As a matter of fact, a high CP accumulation induces RNA encapsidation, preventing its replication by NIb. To enable replication, this excess of CPs is likely to be addressed to ubiquitination-associated degradation, through an interaction with the chaperones CPIP and HSP70 [98,216]. The ELMs analysis of CP IDRs revealed a well-conserved USP7-binding motif (Table S2). As discussed in the case of P1, this motif could participate to CP recognition by a HSP70 dependent ubiquitin-ligase.

Conclusion

Using a proteome dataset representative of the entire *Potyvirus* genus, we were able to analyze the proteome intrinsic disorder both at inter- and intra-species scales. Our work revealed that potyvirus proteomes display a high disorder content. Its maintenance during potyvirus evolution strongly suggests that functional benefits are associated to this structural feature.

A deeper analysis of disorder conservation indicates that, as previously reported for animal viruses, many potyvirus cycle steps and evolutive processes potentially benefit from intrinsic

disorder. This feature could favor potyviral adaptation, as IDRs globally evolve faster than ordered regions, suggesting that they are more tolerant to mutation than structured domains. Our results also suggest that intrinsic disorder regulates the polyprotein proteolytic cleavage. Based on their disorder content and its supposed related functions, potyviral proteins can be mainly classified into three groups. The first group includes P1, VPg and CP, three highly disordered proteins. They all contain IDRs displaying a high aa polymorphism, a distinctive feature of flexible disorder. The functional annotation of these proteins allows to classify them as date hubs, that interact sequentially with several partners. Moreover, their numerous IDRs display several conserved ELM, mostly related to post-translational modifications. This illustrates the potential involvement of intrinsic disorder in host motif mimicry by the virus. The second group, with medium disorder content, includes HC-Pro, P3, P3N-PIPO and CI. Most IDRs of these proteins are rather conserved and belong to the class of constrained disorder. This disorder is especially present in party hubs, which are proteins involved in several simultaneous interactions, as illustrated by P3 scaffolding role in the intracellular virus factories. The third group, represented by two enzymes, NIa-Pro and NIb, contains a low amount of disorder.

The constant progress in our understanding of potyvirus biology underlines its molecular complexity, and many key features of the viral cycle remain unknown. We believe that this proteome-wide analysis of intrinsic disorder provides an alternative way to functionally annotate potyviral proteins, helping to define new experimental paths for exploring the biology of this major viral genus.

Acknowledgments

We thank Dr. Jocelyne Walter, Dr. Benoît Moury, Dr. Eugénie Hébrard and Dr. Thierry Candresse for their helpful comments that led to improve this work. We also thank SharCo consortium for providing us PPV genome sequences used in this work also as Dr. Lukasz Kurgan for his essential help concerning MoRFPred predictions.

References

1 P. E. Wright and H. J. Dyson, *J. Mol. Biol.*, 1999, **293**, 321–31.

2 V. N. Uversky, J. R. Gillespie and A. L. Fink, *Proteins Struct. Funct. Genet.*, 2000, **427**, 415–427.

3 A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner and Z. Obradovic, *J. Mol. Graph. Model.*, 2001, **19**, 26–59.

4 P. Tompa, *Trends Biochem. Sci.*, 2002, **27**, 527–533.

5 K. W. Plaxco and M. Grob, *Nature*, 1997, **386**, 657–658.

6 K. A. Dill and H. S. Chan, *Nat. Perspect.*, 1997, **4**, 10–19.

7 G. W. Daughdrill, M. S. Chadsey, J. E. Karlinsey, T. H. Kelly and F. W. Dahlquist, *Nature*, 1997, **4**, 285–291.

8 R. W. Kriwacki, L. Hengst, L. Tennant, S. I. Reed and P. E. Wright, *Proc Natl Acad Sci USA*, 1996, **93**, 11504–11509.

9 K. Dunker, E. Garner, S. Guilliot, P. Romero, K. Albrecht, Z. Obradovic, C. Science, C. Kissinger, J. E. Villafranca, A. Pharmaceutical and S. Diego, *Pac Symp Biocomput.*, 1998, 473–484.

10 V. N. Uversky, *Biopolymers*, 2013, **99**, 870–87.

11 L. M. Iakoucheva, P. Radivojac, C. J. Brown, T. R. O'Connor, J. G. Sikes, Z. Obradovic and a K. Dunker, *Nucleic Acids Res.*, 2004, **32**, 1037–49.

12 M. O. Collins, L. Yu, I. Campuzano, S. G. N. Grant and J. S. Choudhary, *Mol. Cell. Proteomics*, 2008, **7**, 1331–48.

13 F. Diella, N. Haslam, C. Chica, A. Budd, S. Michael, N. P. Brown, G. Travé and T. J. Gibson, *Front. Biosci.*, 2008, **13**, 6580–6603.

14 C. A. Galea, Y. Wang, S. G. Sivakolundu and R. W. Kriwacki, *Biochemistry*, 2008, **47**, 7598–7609.

15 C. J. Brown, S. Takayama, A. M. Campen, P. Vise, T. W. Marshall, C. J. Oldfield, C. J. Williams and a K. Dunker, *J. Mol. Evol.*, 2002, **55**, 104–10.

16 R. Pancsa and P. Tompa, *PLoS One*, 2012, **7**, e34687.

17 J. W. Chen, P. Romero, V. N. Uversky and A. Keith, *J. Proteome Res.*, 2006, **5**, 888–898.

18 C. J. Brown, A. K. Johnson and G. W. Daughdrill, *Mol. Biol. Evol.*, 2010, **27**, 609–21.

19 J. Bellay, S. Han, M. Michaut, T. Kim, M. Costanzo, B. J. Andrews, C. Boone, G. D. Bader, C. L. Myers and P. M. Kim, *Genome Biol.*, 2011, **12**, R14.

20 C. J. Brown, A. K. Johnson and G. W. Daughdrill, *Mol. Biol. Evol.*, 2011, **29**, 443–443.

21 J. Nilsson, M. Grahn and A. P. H. Wright, *Genome Biol.*, 2011, **12**, R65.

22 Z. Peng, J. Yan, X. Fan, M. J. Mizianty, B. Xue, K. Wang, G. Hu, V. N. Uversky and L. Kurgan, *Cell. Mol. Life Sci.*, 2014.

23 B. Xue, A. K. Dunker and V. N. Uversky, *J. Biomol. Struct. Dyn.*, 2012, **30**, 137–149.

24 M. Sickmeier, J. a Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic and a K. Dunker, *Nucleic Acids Res.*, 2007, **35**, D786–93.

25 R. Pushker, C. Mooney, N. E. Davey, J.-M. Jacqué and D. C. Shields, *PLoS One*, 2013, **8**, e60724.

26 V. N. Uversky and S. Longhi, *Flexible Viruses : Structural disorder in Viral Proteins*, John Wiley and Sons : Hoboken, 2012.

27 B. Xue, D. Blocquel, J. Habchi, A. V Uversky, L. Kurgan, V. N. Uversky and S. Longhi, *Chem. Rev.*, 2014, **114**, 6880–911.

28 C. Alves and C. Cunha, *Future Virol.*, 2012, **7**, 1183–1191.

29 E. Hébrard, Y. Bessin, T. Michon, S. Longhi, V. N. Uversky, F. Delalande, A. Van Dorsselaer, P. Romero, J. Walter, N. Declerck and D. Fargette, *Virol. J.*, 2009, **6**, 23.

30 N. Tokuriki, C. J. Oldfield, V. N. Uversky, I. N. Berezovsky and D. S. Tawfik, *Trends Biochem. Sci.*, 2009, **34**, 53–9.

31 S. Duffy, L. A. Shackelton and E. C. Holmes, *Nat. Rev. Genet.*, 2008, **9**, 267–76.

32 G. K.-M. Goh, A. K. Dunker and V. N. Uversky, *BMC Genomics*, 2008, **9**, S4.

33 B. Xue and V. N. Uversky, *J. Mol. Biol.*, 2013, **426**, 1322–1350.

34 G. K.-M. Goh, A. K. Dunker and V. N. Uversky, *J. Pathog.*, 2012, **2012**, 738590.

35 G. K.-M. Goh, a K. Dunker and V. N. Uversky, *Virol. J.*, 2009, **6**, 69.

36 G. K.-M. Goh, A. K. Dunker and V. N. Uversky, *Plos Curr. outbreaks*, 2013.

37 T. Hagai, A. Azia, M. M. Babu and R. Andino, *Cell Rep.*, 2014, **7**, 1729–1739.

38 L. Gitlin, T. Hagai, A. LaBarbera, M. Solovey and R. Andino, *PLoS Pathog.*, 2014, **10**, e1004529.

39 X. Fan, B. Xue, P. T. Dolan, D. J. Lacount, L. Kurgan and V. N. Uversky, *Mol. Biosyst.*, 2014, **10**, 1345–1363.

40 B. Xue, M. J. Mizianty, L. Kurgan and V. N. Uversky, *Cell. Mol. Life Sci.*, 2012, **69**, 1211–59.

41 V. N. Uversky, A. Roman, C. J. Oldfield and A. K. Dunker, *J. Proteome Res.*, 2006, **5**, 1829–1842.

42 F. Meng, R. A. Badierah, H. A. Almehdar, E. M. Redwan, L. Kurgan and V. N. Uversky, *FEBS J.*, 2015, **282**, 3368–94.

43 J. F. Ortiz, M. L. MacDonald, P. Masterson, V. N. Uversky and J. Siltberg-Liberles, *Genome Biol. Evol.*, 2013, **5**, 504–13.

44 J. Chroboczek, E. Hébrard, K. Mäkinen, T. Michon and K. Rantalainen, in *Flexible Viruses*, 2012.

45 R. Grzela, E. Szolajska, C. Ebel, D. Madern, A. Favier, I. Wojtal, W. Zagorski and J. Chroboczek, *J. Biol. Chem.*, 2008, **283**, 213–21.

46 K. I. Rantalainen, V. N. Uversky, P. Permi, N. Kalkkinen, A. keith Dunker and K. Makinen, *Virology*, 2008, **377**, 280–288.

47 K. I. Rantalainen, K. Eskelin, P. Tompa and K. Mäkinen, *J. Virol.*, 2011, **85**, 2449–57.

48 A. L. Ksenofontov, V. Paalme, A. M. Arutyunyan, P. I. Semenyuk, N. V Fedorova, R. Rumvolt, L. A. Baratova, L. Järvekülg and E. N. Dobrov, *PLoS One*, 2013, **8**, e67830.

49 V. Nicaise, *Front. Plant Sci.*, 2014, **5**, 1–18.

50 C. Ward and D. Shukla, *Intervirology*, 1991, **32**, 269–296.

51 K.-B. G. Scholthof, S. Adkins, H. Czosnek, P. Palukaitis, E. Jacquot, T. Hohn, B. Hohn, K. Saunders, T. Candresse, P. Ahlquist, C. Hemenway and G. D. Foster, *Mol. Plant Pathol.*, 2011, **12**, 938–54.

52 D. Shukla and C. Ward, *Adv. Virus Res.*, 1989, **36**, 273–314.

53 M. J. Adams, F. . Zerbini, R. French, F. Rabenstein, D. . Stenger and J. P. T. Valkonen, *ICTV*, 2012, 1069–1089.

54 A. Kendall, M. McDonald, W. Bian, T. Bowles, S. C. Baumgarten, J. Shi, P. L. Stewart, E. Bullitt, D. Gore, T. C. Irving, W. M. Havens, S. a Ghabrial, J. S. Wall and G. Stubbs, *J. Virol.*, 2008, **82**, 9546–54.

55 S. Lain, J. L. Riechmann, E. Mcndez and J. A. Garcia, *Virus Res.*, 1988, **10**.

56 M. F. E. Siaw, M. Shahabuddin, S. Ballard, J. G. Shaw and R. E. Rhoads, *Virology*, 1985, **142**, 134–143.

57 J. L. Riechmann, S. Lain and J. A. Garcia, *J Gen Virol*, 1989, **70**, 2785–2789.

58 A. Merits, M. Rajama, P. Runeberg-roos and T. Kekarainen, *J. Gen. Virol.*, 2002, 1211–1221.

59 J. C. Carrington, D. D. Freed and C.-S. Oh, *EMBO J.*, 1990, **9**, 1347–1353.

60 J. C. Carrington, S. M. Cary, T. D. Parks and W. G. Dougherty, *EMBO J.*, 1989, **8**, 365–370.

61 J. Verchot, E. V. Koonin and J. C. Carrington, *Virology*, 1991, **185**, 527–535.

62 J. A. Garcia, J. L. Riechmann and S. Lain, *J. Virol.*, 1989, **63**, 2457–2460.

63 B. Y.-W. Chung, W. A. Miller, J. F. Atkins and A. E. Firth, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 5897–902.

64 Y. Hong and A. G. Hunt, *Virology*, 1996, **226**, 146–151.

65 K. Mäkinen and A. Hafrén, *Front. Plant Sci.*, 2014, **5**, 1–12.

66 K. Tamura, G. Stecher, D. Peterson, A. Filipski and S. Kumar, *Mol. Biol. Evol.*, 2013, **30**, 2725–9.

67 J. Hillung, S. F. Elena and J. M. Cuevas, *BMC Evol. Biol.*, 2013, **13**, 249.

68 P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown and A. K. Dunker, *Proteins Struct. Funct. Genet.*, 2001, **48**, 38–48.

69 F.-C. Chen, C.-L. Pan and H.-Y. Lin, *Mol. Biol. Evol.*, 2012, **29**, 187–93.

70 K. Scheffler, D. P. Martin and C. Seoighe, *Bioinformatics*, 2006, **22**, 2493–9.

71 S. L. K. Pond, S. D. W. Frost and S. V Muse, *Bioinformatics*, 2005, **21**, 676–9.

72 W. Delport, A. F. Y. Poon, S. D. W. Frost and S. L. Kosakovsky Pond, *Bioinformatics*, 2010, **26**, 2455–7.

73 S. L. K. Pond and S. D. W. Frost, *Bioinformatics*, 2005, **21**, 2531–3.

74 S. L. K. Pond, B. Murrell and A. F. Y. Poon, *Methods Mol. Biol.*, 2012, **856**, 239–72.

75 M. Anisimova, R. Nielsen and Z. Yang, *Genetics*, 2003, **1236**, 1229–1236.

76 K. Katoh, K. Kuma, H. Toh and T. Miyata, *Nucleic Acids Res.*, 2005, **33**, 511–8.

77 A. Gibbs and K. Ohshima, *Annu. Rev. Phytopathol.*, 2010, **48**, 205–223.

78 M. Garcia-Boronat, C. M. Diez-Rivero, E. L. Reinherz and P. a Reche, *Nucleic Acids Res.*, 2008, **36**, W35–41.

79 F. M. Disfani, W.-L. Hsu, M. J. Mizianty, C. J. Oldfield, B. Xue, a K. Dunker, V. N. Uversky and L. Kurgan, *Bioinformatics*, 2012, **28**, 75–83.

80 N. E. Davey, K. Van Roey, R. J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel and T. J. Gibson, *Mol. Biosyst.*, 2012, **8**, 268–81.

81 A. J. Gibbs, K. Ohshima, M. J. Phillips and M. J. Gibbs, *PLoS One*, 2008, **3**, e2523.

82 K. Rajagopalan, S. M. Mooney, N. Parekh, R. H. Getzenberg and P. Kulkarni, *J. Cell. Biochem.*, 2011, **112**, 3256–67.

83 C. Haynes, C. J. Oldfield, F. Ji, N. Klitgord, M. E. Cusick, P. Radivojac, V. N. Uversky, M. Vidal and L. M. Iakoucheva, *PLoS Comput Biol*, 2006, **2**, e100.

84 P. M. Kim, A. Sboner, Y. Xia and M. Gerstein, *Mol. Syst. Biol.*, 2008, **4**, 179.

85 L. C. James and D. S. Tawfik, *Trends Biochem. Sci.*, 2003, **28**, 361–8.

86 S. F. Elena and G. Rodrigo, *Curr. Opin. Virol.*, 2012, **2**, 719–24.

87 F. Revers and J. García, in *Advances in Virus Research, Vol.92*, 2015, pp. 101–199.

88 M. J. Adams, J. F. Antoniw and F. Beaudoin, *Mol. Plant Pathol.*, 2005, **6**, 471–87.

89 M. A. Restrepo-hartwig and J. C. Carrington, *J. Virol.*, 1994, **68**, 2388–2397.

90 M. C. Schaad, P. E. Jensen and J. C. Carrington, *EMBO J.*, 1997, **16**, 4049–59.

91 S. Leonard, *J. Gen. Virol.*, 2004, **85**, 1055–1063.

92 E. Majer, Z. Salvador, M. P. Zwart, A. Willemsen, S. F. Elena and J.-A. Daròs, *J. Virol.*, 2014, **88**, 4586–90.

93 C. Mathur and H. S. Savithri, *Biochem. Biophys. Res. Commun.*, 2012, **427**, 113–8.

94 A. Fontana, P. P. De Laureto, B. Spolaore, E. Frare, P. Picotti and M. Zambonin, *Acta Biochim. Pol.*, 2004, **51**, 299–321.

95 D. E. Johnson, B. Xue, M. D. Sickmeier, J. Meng, M. S. Cortese, C. J. Oldfield, T. Le Gall, A. K. Dunker and V. N. Uversky, *J. Struct. Biol.*, 2012, **180**, 201–15.

96 P. Paulverino De Laureto, V. de Filippis, E. Scaramella, M. Zambonin and A. Fontana, *Eur. J. Biochem.*, 1995, **230**, 779–787.

97 P. Paulverino De Laureto, L. Tosatto, E. Frare, O. Marin, V. N. Uversky and A. Fontana, *Biochemistry*, 2006, **45**, 11523–11531.

98 J. Besong-Ndika, K. I. Ivanov, A. Hafrèn, T. Michon and K. Mäkinen, *J. Virol.*, 2015, **89**, 4237–48.

99 G. W. Daughdrill, P. Narayanaswami, S. H. Gilmore, A. Belczyk and C. J. Brown, *J. Mol. Evol.*, 2007, **65**, 277–288.

100 J. J. Han, N. Bertin, T. Hao and D. S. Goldberg, *Nature*, 2004, **430**, 88–93.

101 R. Van Der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Old, R. V Pappu, P. Tompa, V. N. Uversky, P. E. Wright and M. M. Babu, *Chem. Rev.*, 2014, **114**, 6589–6631.

102 P. M. Kim, L. J. Lu, Y. Xia and B. G. Mark, *Science (80-.).*, 2006, **603**.

103 S. Agarwal, C. M. Deane, M. a Porter and N. S. Jones, *PLoS Comput. Biol.*, 2010, **6**, e1000817.

104 C. J. Oldfield, Y. Cheng, M. S. Cortese, P. Romero, V. N. Uversky and A. K. Dunker, *Biochemistry*, 2005, **44**, 12454–70.

105 A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, a K. Dunker and V. N. Uversky, *J. Mol. Biol.*, 2006, **362**, 1043–59.

106 P. T. Dolan, A. P. Roth, B. Xue, R. Sun, a K. Dunker, V. N. Uversky and D. J. LaCount, *Protein Sci.*, 2015, **24**, 221–35.

107 B. Meszaros, I. Simon and Z. Dosztanyi, *PLoS Comput. Biol.*, 2009, **5**, e1000376.

108 V. Vacic, C. J. Oldfield, A. Mohan, P. Radivojac, M. S. Cortese, V. N. Uversky and A. K. Dunker, *J. Proteome Res.*, 2007, **6**, 2351–2366.

109 H. Dinkel, K. Van Roey, S. Michael, N. E. Davey, R. J. Weatheritt, D. Born, T. Speck, D. Krüger, G. Grebnev, M. Kuban, M. Strumillo, B. Uyar, A. Budd, B. Altenberg, M. Seiler, L. B. Chemes, J. Glavina, I. E. Sánchez, F. Diella and T. J. Gibson, *Nucleic Acids Res.*, 2014, **42**, D259–66.

110 H. Dinkel, S. Michael, R. J. Weatheritt, N. E. Davey, K. Van Roey, B. Altenberg, G. Toedt, B. Uyar, M. Seiler, A. Budd, L. Jödicke, M. a Dammert, C. Schroeter, M. Hammer, T. Schmidt, P. Jehl, C. McGuigan, M. Dymecka, C. Chica, K. Luck, A. Via, A. Chatr-Aryamontri, N. Haslam, G. Grebnev, R. J. Edwards, M. O. Steinmetz, H. Meiselbach, F. Diella and T. J. Gibson, *Nucleic Acids Res.*, 2012, **40**, D242–51.

111 M. Fuxreiter, P. Tompa and I. Simon, *Bioinformatics*, 2007, **23**, 950–6.

112 R. Pancsa and M. Fuxreiter, *IUBMB Life*, 2012, **64**, 513–520.

113 N. E. Davey, G. Travé and T. J. Gibson, *Trends Biochem. Sci.*, 2011, **36**, 159–69.

114 N. C. Elde and H. S. Malik, *Nat. Rev. Microbiol.*, 2009, **7**, 787–97.

115 J. Verchot and J. C. Carrington, *J Virol*, 1995, **69**, 3668–74.

116 A. Valli, J. J. López-Moya and J. A. García, *J. Gen. Virol.*, 2007, **88**, 1016–28.

117 J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton and D. T. Jones, *J. Mol. Biol.*, 2004, **337**, 635–45.

118 H. J. Dyson, *Mol. Biosyst.*, 2012, **8**, 97–104.

119 Y. Soumounou and J.-F. Laliberté, *J. Gen. Virol.*, 1994, **75**, 2567–2573.

120 F. Martinez and J. A. Daros, *J Virol*, 2014, **88**, 10725–10737.

121 P. Tompa and P. Csermely, *FASEB J.*, 2004, **18**, 1169–75.

122 R. Ivanyi-Nagy, J. P. Lavergne, C. Gabus, D. Ficheux and J. L. Darlix, *Nucleic Acids Res.*, 2008, **36**, 712–725.

123 F. Pasin, C. Simón-Mateo and J. A. García, *PLoS Pathog.*, 2014, **10**, e1003985.

124 T. Trudeau, R. Nassar, A. Cumberworth, E. T. C. Wong, G. Woollard and J. Gsponer, *Structure*, 2013, **21**, 332–341.

125 A. Merits, D. Guo, L. Järvekülg and M. Saarma, *Virology*, 1999, **263**, 15–22.

126 V. I. Maliogka, B. Salvador, A. Carbonell, P. Sáenz, D. S. León, J. C. Oliveros, M. O. Delgadillo, J. A. García and C. Simón-Mateo, *Mol. Plant Pathol.*, 2012, **13**, 877–86.

127 B. Salvador, P. Saénz, E. Yangüez, J. B. Quiot, L. Quiot, M. O. Delgadillo, J. A. García and C. Simón-Mateo, *Mol. Plant Pathol.*, 2008, **9**, 147–55.

128 C. Brown and A. Johnson, *Curr. Opin. Struct. Biol.*, 2011, **21**, 441–446.

129 A. Schlessinger, C. Schaefer, E. Vicedo, M. Schmidberger, M. Punta and B. Rost, *Curr. Opin. Struct. Biol.*, 2011, **21**, 412–8.

130 R. Lozano-Durán and S. Robatzek, *Mol. Plant-Microbe Interact.*, 2015, **28**, 511–8.

131 D. Chevalier, E. R. Morris and J. C. Walker, *Annu. Rev. Plant Biol.*, 2009, **60**, 67–91.

132 J. L. Carrasco, M. J. Castelló, K. Naumann, I. Lassowskat, M. Navarrete-Gómez, D. Scheel and P. Vera, *PLoS One*, 2014, **9**, e90734.

133 Z. Ding, G. I. Lee, X. Liang, F. Gallazzi, a. Arunima and S. R. Van Doren, *Biochemistry*, 2005, **44**, 10119–10134.

134 J. Qin, D. Barajas and P. D. Nagy, *Virology*, 2012, **426**, 106–19.

135 C. Alcaide-Loridan and I. Jupin, *Plant Physiol.*, 2012, **160**, 72–82.

136 M. Chenon, L. Camborde, S. Cheminant and I. Jupin, *EMBO J.*, 2012, **31**, 741–753.

137 S. E. Dixon, M. M. Bhatti, V. N. Uversky, K. Dunker and W. J. Sullivan Jr, *Mol Biochem Parasitol*, 2011, **175**, 192–195.

138 F. Martínez and J.-A. Daròs, *J. Virol.*, 2014, **88**, 10725–37.

139 J. Syller, *Physiol. Mol. Plant Pathol.*, 2006, **67**, 119–130.

140 C. Plisson, M. Drucker, S. Blanc, S. German-Retana, O. Le Gall, D. Thomas and P. Bron, *J. Biol. Chem.*, 2003, **278**, 23753–61.

141 Y. M. Shiboleth, E. Haronsky, D. Leibman, T. Arazi, M. Wassenegger, S. A. Whitham, V. Gaba and A. Gal-On, *J. Virol.*, 2007, **81**, 13135–13148.

142 R. M. Jamous, K. Boonrod, M. W. Fuellgrabe, M. S. Ali-Shtayeh, G. Krczal and M. Wassenegger, *J. Gen. Virol.*, 2011, **92**, 2222–6.

143 B. Guo, J. Lin and K. Ye, *J. Biol. Chem.*, 2011, **286**, 21937–21943.

144 N. Suehiro, T. Natsuaki, T. Watanabe and S. Okuda, *J. Gen. Virol.*, 2004, **85**, 2087–2098.

145 X. Cui, T. Wei, R. V Chowda-Reddy, G. Sun and A. Wang, *Virology*, 2010, **397**, 56–63.

146 S. Eiamtanasate, M. Juricek and Y.-K. Yap, *Virus Genes*, 2007, **35**, 611–7.

147 A. B. Sigalov and V. N. Uversky, *Self. Nonself.*, 2011, **2**, 55–72.

148 L. Lin, Z. Luo, F. Yan, Y. Lu, H. Zheng and J. Chen, *Virus Genes*, 2011, **43**, 90–2.

149 T. Wei, C. Zhang, J. Hong, R. Xiong, K. D. Kasschau, X. Zhou, J. C. Carrington and A. Wang, *PLoS Pathog.*, 2010, **6**, e1000962.

150 R.-H. Wen and M. R. Hajimorad, *Virology*, 2010, **400**, 1–7.

151 P. Vijayapalani, M. Maeshima, N. Nagasaki-Takekuchi and W. A. Miller, *PLoS Pathog.*, 2012, **8**, e1002639.

152 C. Rancurel, M. Khosravi, a K. Dunker, P. R. Romero and D. Karlin, *J. Virol.*, 2009, **83**, 10719–36.

153 M. Sorel, J. a Garcia and S. German-Retana, *Mol. plant-microbe Interact.*, 2014, **27**, 215–26.

154 A. Fernández, S. Laín and J. García, *Nucleic Acids Res.*, 1995, **23**, 1327–1332.

155 G. Tavert-Roudet, A. Abdul-Razzak, B. Doublet, J. Walter, T. Delaunay, S. German-Retana, T. Michon, O. Le Gall and T. Candresse, *J. Gen. Virol.*, 2012, **93**, 184–93.

156 M. E. Fairman-Williams, U.-P. Guenther and E. Jankowsky, *Curr. Opin. Struct. Biol.*, 2010, **20**, 313–24.

157 V. N. Uversky, *Intrinsically Disord. Proteins*, 2014, **2**, 1–42.

158 I. Oruetxebarria, D. Guo, A. Merits and K. Ma, *Virus Res.*, 2001, **73**, 103–112.

159 R. Anindya, S. Chittori and H. S. Savithri, *Virology*, 2005, **336**, 154–62.

160 K. Eskelin, A. Hafrén, K. I. Rantalainen and K. Mäkinen, *J. Virol.*, 2011, **85**, 9210–21.

161 M. a Khan, H. Miyoshi, S. Ray, T. Natsuaki, N. Suehiro and D. J. Goss, *J. Biol. Chem.*, 2006, **281**, 28002–10.

162 C. Robaglia and C. Caranta, *Trends Plant Sci.*, 2006, **11**, 40–5.

163 M. a Khan, H. Miyoshi, D. R. Gallie and D. J. Goss, *J. Biol. Chem.*, 2008, **283**, 1340–9.

164 H. Miyoshi, H. Okade, S. Muto, N. Suehiro, H. Nakashima, K. Tomoo and T. Natsuaki, *Biochimie*, 2008, **90**, 1427–1434.

165 T. Michon, Y. Estevez, J. Walter, S. German-Retana and O. Le Gall, *FEBS J.*, 2006, **273**, 1312–22.

166 R. Gabrenaite-Verkhovskaya, I. a Andreev, N. O. Kalinina, L. Torrance, M. E. Taliansky and K. Mäkinen, *J. Gen. Virol.*, 2008, **89**, 829–38.

167 M.-L. Rajamäki and J. P. T. Valkonen, *Mol. Plant. Microbe. Interact.*, 2003, **16**, 25–34.

168 Z. Gao, E. Johansen, S. Eyers, C. L. Thomas, T. H. Noel Ellis and A. J. Maule, *Plant J.*, 2004, **40**, 376–85.

169 P. Dunoyer, C. Thomas, S. Harrison, F. Revers and A. Maule, 2004, **78**, 2301–2309.

170 J. Herold and R. Andino, *Mol. Cell*, 2001, **7**, 581–591.

171 P. Puustinen and K. Mäkinen, *J. Biol. Chem.*, 2004, **279**, 38103–10.

172 J. Jiang and J.-F. Laliberté, *Curr. Opin. Virol.*, 2011, **1**, 347–54.

173 S. Wittmann, H. Chatel, M. G. Fortin and J. F. Laliberté, *Virology*, 1997, **234**, 84–92.

174 S. Léonard, D. Plante, S. Wittmann, N. Daigneault, M. G. Fortin and J. F. Laliberté, *J. Virol.*, 2000, **74**, 7730–7.

175 M. C. Schaad, R. J. Anderberg and J. C. Carrington, *Virology*, 2000, **273**, 300–6.

176 C. Beauchemin, N. Boutet and J.-F. Laliberté, *J. Virol.*, 2007, **81**, 775–82.

177 C. Charron, M. Nicolaï, J.-L. Gallois, C. Robaglia, B. Moury, A. Palloix and C. Caranta, *Plant J.*, 2008, **54**, 56–68.

178 B. Culjkovic, I. Topisirovic and K. L. B. Borden, *Cell Cycle*, 2007, **6**, 65–69.

179 C. L. Bokros, J. D. Hugdahl, H. Kim, V. R. Hanesworth, A. Van Heerden, K. S. Browning and L. C. Morejohn, *Proc. Natl. Acad. Sci. U. S. A.*, 1995, **92**, 7120–7124.

180 M.-L. Rajamäki and J. P. T. Valkonen, *Plant Cell*, 2009, **21**, 2485–502.

181 T.-S. Huang, T. Wei, J.-F. Laliberté and A. Wang, *Plant Physiol.*, 2010, **152**, 255–66.

182 K. Thivierge, S. Cotton, P. J. Dufresne, I. Mathieu, C. Beauchemin, C. Ide, M. G. Fortin and J.-F. Laliberté, *Virology*, 2008, **377**, 216–25.

183 G. Roudet-Tavert, T. Michon, J. Walter, T. Delaunay, E. Redondo and O. Le Gall, *J. Gen. Virol.*, 2007, **88**, 1029–33.

184 B. Moury, C. Charron, B. Janzac, V. Simon, J. L. Gallois, A. Palloix and C. Caranta, *Infect. Genet. Evol.*, 2014, **27**, 472–80.

185 V. Ayme, S. Souche, C. Caranta, M. Jacquemond, J. Chadoeuf, A. Palloix and B. Moury, *Mol. Plant. Microbe. Interact.*, 2006, **19**, 557–63.

186 B. Moury, C. Morel, E. Johansen, L. Guilbaud, S. Souche, V. Ayme, C. Caranta, A. Palloix and M. Jacquemond, *Mol. Plant. Microbe. Interact.*, 2004, **17**, 322–9.

187 M. Bruun-Rasmussen, I. S. Møller, G. Tulinius, J. K. R. Hansen, O. S. Lund and I. E. Johansen, *Mol. Plant-Microbe Interact.*, 2007, **20**, 1075–1082.

188 P. Puustinen, M. Rajama, K. I. Ivanov and J. P. T. Valkonen, *J. Virol.*, 2002, **76**, 12703–12711.

189 A. Hafrén and K. Mäkinen, *J. Gen. Virol.*, 2008, **89**, 1509–18.

190 J. C. Carrington, D. D. Freed and A. J. Leinicke, *Plant Cell*, 1991, **3**, 953–962.

191 R. Anindya and H. S. Savithri, *J. Biol. Chem.*, 2004, **279**, 32159–69.

192 J. Phan, A. Zdanov, A. G. Evdokimov, J. E. Tropea, H. K. Peters, R. B. Kapust, M. Li, A. Wlodawer and D. S. Waugh, *J. Biol. Chem.*, 2002, **277**, 50564–72.

193 P. Sun, B. P. Austin, J. Tözsér and D. S. Waugh, *Protein Sci.*, 2010, **19**, 2240–51.

194 C. M. Nunn, M. Jeeves, M. J. Cliff, G. T. Urquhart, R. R. George, L. H. Chao, Y. Tscuchia and S. Djordjevic, *J. Mol. Biol.*, 2005, **350**, 145–155.

195 Y. Wang, C. Liu, D. Yang, H. Yu and Y.-C. Liou, *Mol. Cell*, 2010, **37**, 112–22.

196 P. J. Dufresne, K. Thivierge, S. Cotton, C. Beauchemin, C. Ide, E. Ubalijoro, J.-F. Laliberté and M. G. Fortin, *Virology*, 2008, **374**, 217–27.

197 X. Wang, Z. Ullah and R. Grumet, *Virology*, 2000, **275**, 433–43.

198 L. Liljas, in *Flexible Viruses*, eds. S. Longhi and V. N. Uversky, John Wiley & Sons, Inc;, 2012, pp. 35–46.

199 K. I. Ivanov and K. Mäkinen, *Curr. Opin. Virol.*, 2012, **2**, 712–8.

200 P. Ni and K. C. Cheng, *Virology*, 2013, **446**, 123–32.

201 V. Dolja, R. Haldeman-Cahill, A. Montgomery, K. Vandenbosch and J. Carrington, *Virology*, 1995, **206**, 1007–1016.

202 R. Anindya and H. S. Savithri, *Virology*, 2003, **316**, 325–336.

203 D. D. Shukla, P. M. Strike, S. L. Tracy, K. H. Gough and C. W. Ward, *J Gen Virol*, 1988, **69**, 1497–1508.

204 L. A. Baratova, A. V Efimov, N. Eugenie, N. V Fedorova, R. Hunt, G. a Badun, A. L. Ksenofontov, L. Torrance and E. N. Dobrov, *J. Virol.*, 2001, **75**, 9696–9702.

205 A. . Voloudakis, C. . Malpica, M.-E. Aleman-Verdaguer, D. . Stark, C. . Fauquet and R. . Beachy, *Arch. Virol.*, 2004, **149**, 699–712.

206 P. L. Atreya, C. D. Atreya and T. P. Pirone, *Proc Natl Acad Sci USA*, 1991, **88**, 7887–7891.

207 S. Blanc, J.-J. López-Moya, R. Wang, S. García-Lampasona, D. W. Thornbury and T. P. Pirone, *Virology*, 1997, **231**, 141–147.

208 B. Shoemaker, J. J. Portman and P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 8868–8873.

209 Y. Huang and Z. Liu, *J. Mol. Biol.*, 2009, **393**, 1143–1159.

210 R. F. Allison, W. G. Dougherty, T. D. Parks, L. Willis, R. E. Johnston, M. Kelly and F. B. Armstrong, *Virology*, 1985, **147**, 309–316.

211 C. Desbiez, C. Chandeysson and H. Lecoq, *Mol. Plant Pathol.*, 2014, **15**, 217–21.

212 K. I. Ivanov, P. Puustinen, A. Merits, M. Saarma and K. Mäkinen, *J. Biol. Chem.*, 2001, **276**, 13530–40.

213 K. I. Ivanov, P. Puustinen, R. Gabrenaite, H. Vihinen, L. Rönnstrand, L. Valmu, N. Kalkkinen and K. Mäkinen, *Plant Cell*, 2003, **15**, 2124–2139.

214 Z. W. Subr, M. Kamencayová, S. Nováková, a Nagyová, J. Nosek and M. Glasa, *Arch. Virol.*, 2010, **155**, 1151–5.

215 D. Chen, S. Jua, L. Hartweck, J. M. Alamillo, C. Simo and J. J. Pe, *J. Virol.*, 2005, **79**, 9381–9387.

216 A. Hafrén, D. Hofius, G. Rönnholm, U. Sonnewald and K. Mäkinen, *Plant Cell*, 2010, **22**, 523–35.

Charon et al. Figures

Figures

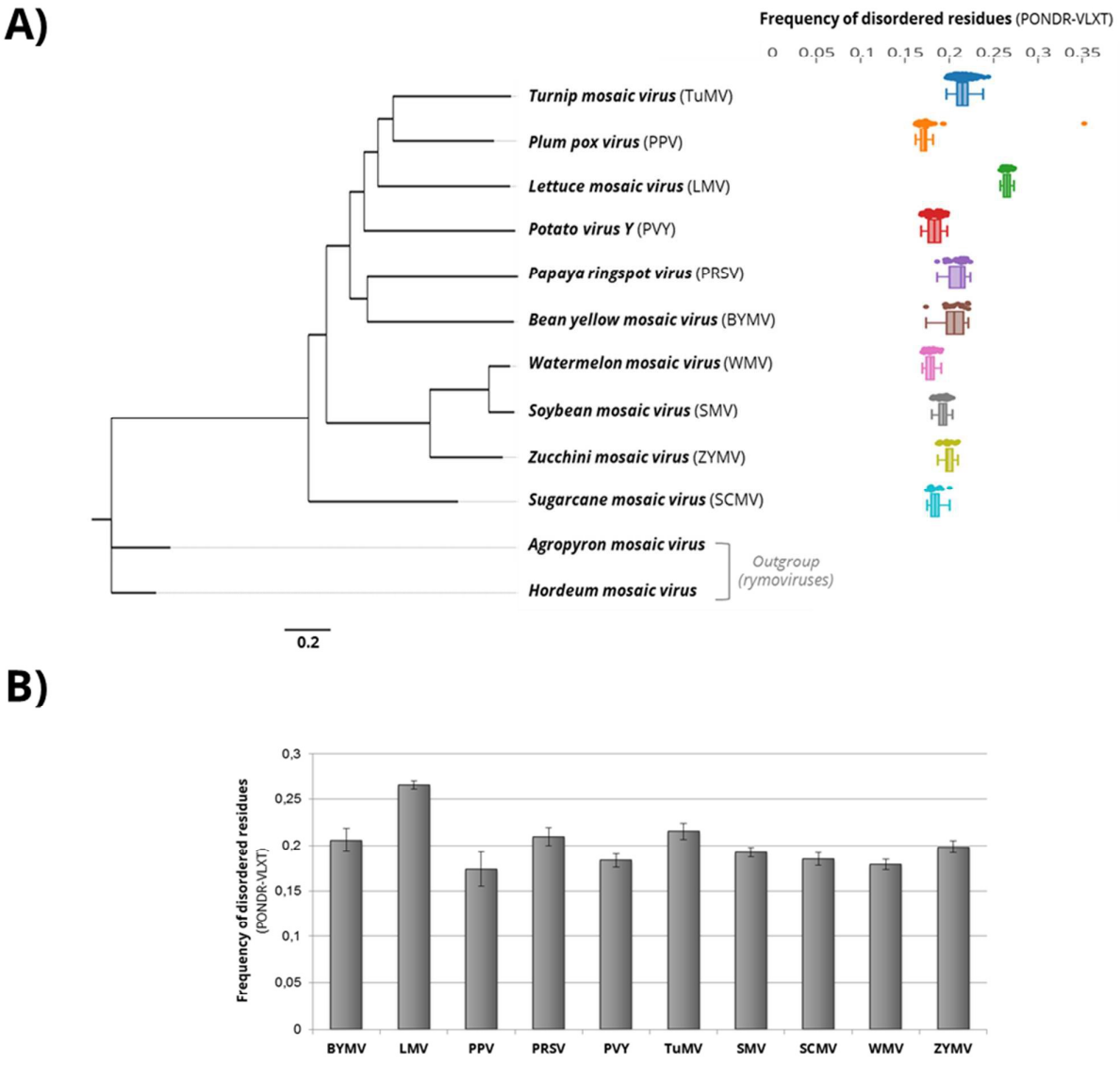


FIG 1. Intrinsic disorder conservation in proteome of 10 potyvirus species. A) Phylogenetic tree of potyvirus polyproteins used in this study. Box plots represent variation of residues frequencies predicted as disordered (PONDR-VLXT) among each potyvirus polyprotein dataset. Dots represent disorder frequency calculated for each polyprotein of the dataset used to construct the plot. Tree scale bar correspond to number of substitutions per site. **B)** Bar charts representation of mean disordered residues frequency for each potyvirus proteome (BYMV, *Bean yellow mosaic virus;* LMV, *Lettuce mosaic virus;* PPV, *Plum pox virus;* PRSV, *Papaya ringspot virus;* PVY, *Potato Virus Y;* TuMV, *Turnip mosaic virus;* SMV, *Soybean mosaic virus;* SCMV, *Sugarcane mosaic virus;* WMV, *Watermelon mosaic virus;* ZYMV, *Zucchini yellow mosaic virus*).

Charon et al. Figures

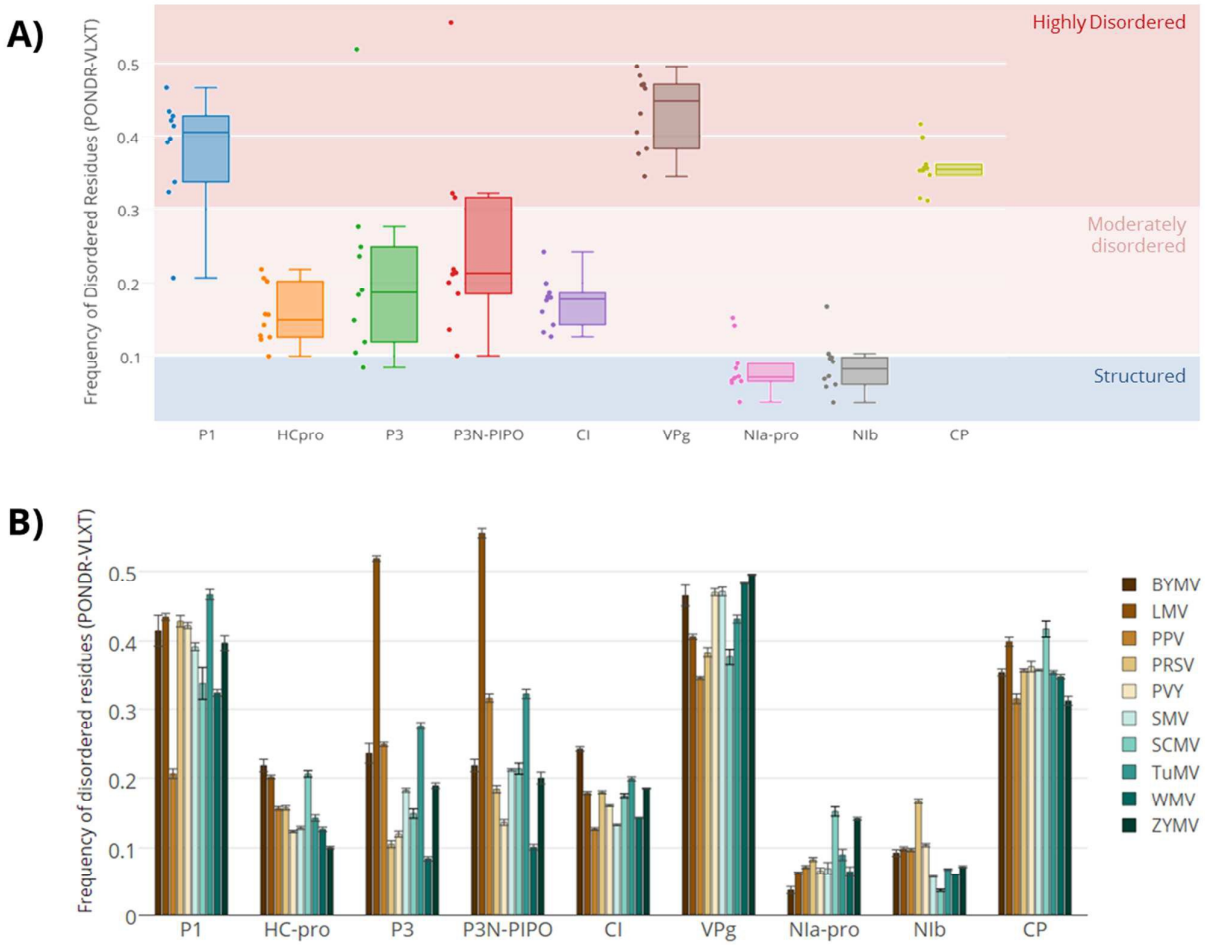


FIG 2. Conservation of intrinsic disorder amount in potyvirus proteins. A) Box plot representation of mean variation in disordered residues frequency (predicted with PONDR-VLXT) for each protein per potyvirus species. Proteins are classified depending on their total disorder content : 0-10% are considered as highly ordered (blue), 11-30% as moderately disordered (white) and 31-100% as highly disordered (red). **B)** Bar charts representation of mean disordered residues frequency for each proteins of each species. (BYMV, *Bean yellow mosaic virus;* LMV, *Lettuce mosaic virus;* PPV, *Plum pox virus;* PRSV, *Papaya ringspot virus;* PVY, *Potato Virus Y;* TuMV, *Turnip mosaic virus;* SMV, *Soybean mosaic virus;* SCMV, *Sugarcane mosaic virus;* WMV, *Watermelon mosaic virus;* ZYMV, *Zucchini mosaic virus*).

Charon et al. Figures

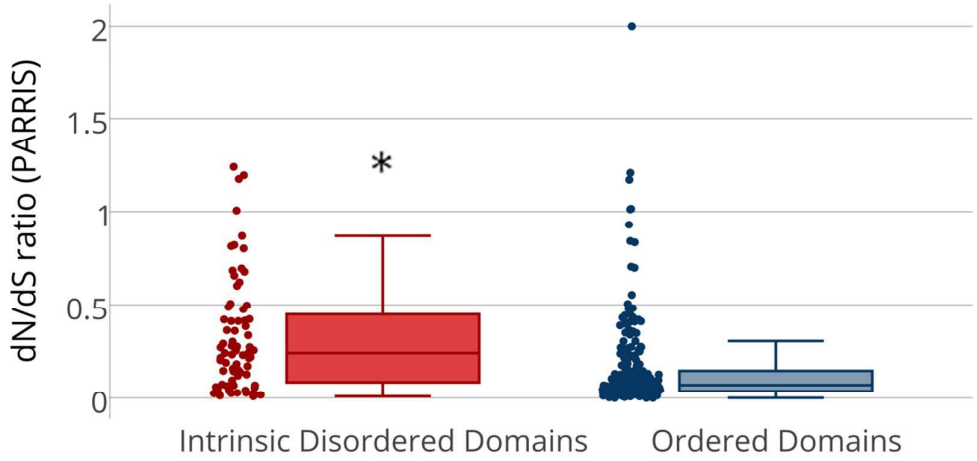


FIG 3. Evolutive constraints (dN/dS ratio) exerted on intrinsic disordered domains and Ordered domains of Potyvirus proteins. Box blots are constructed respectively with mean dN/dS data obtained both for IDRs (red box plot and dots) and ordered domains or OD (blue ones). Dots represent each dN/dS value obtained. Star represent statistical significance of difference between mean OD and IDR dN/dS (p-value<0.001).

Charon et al. Figures

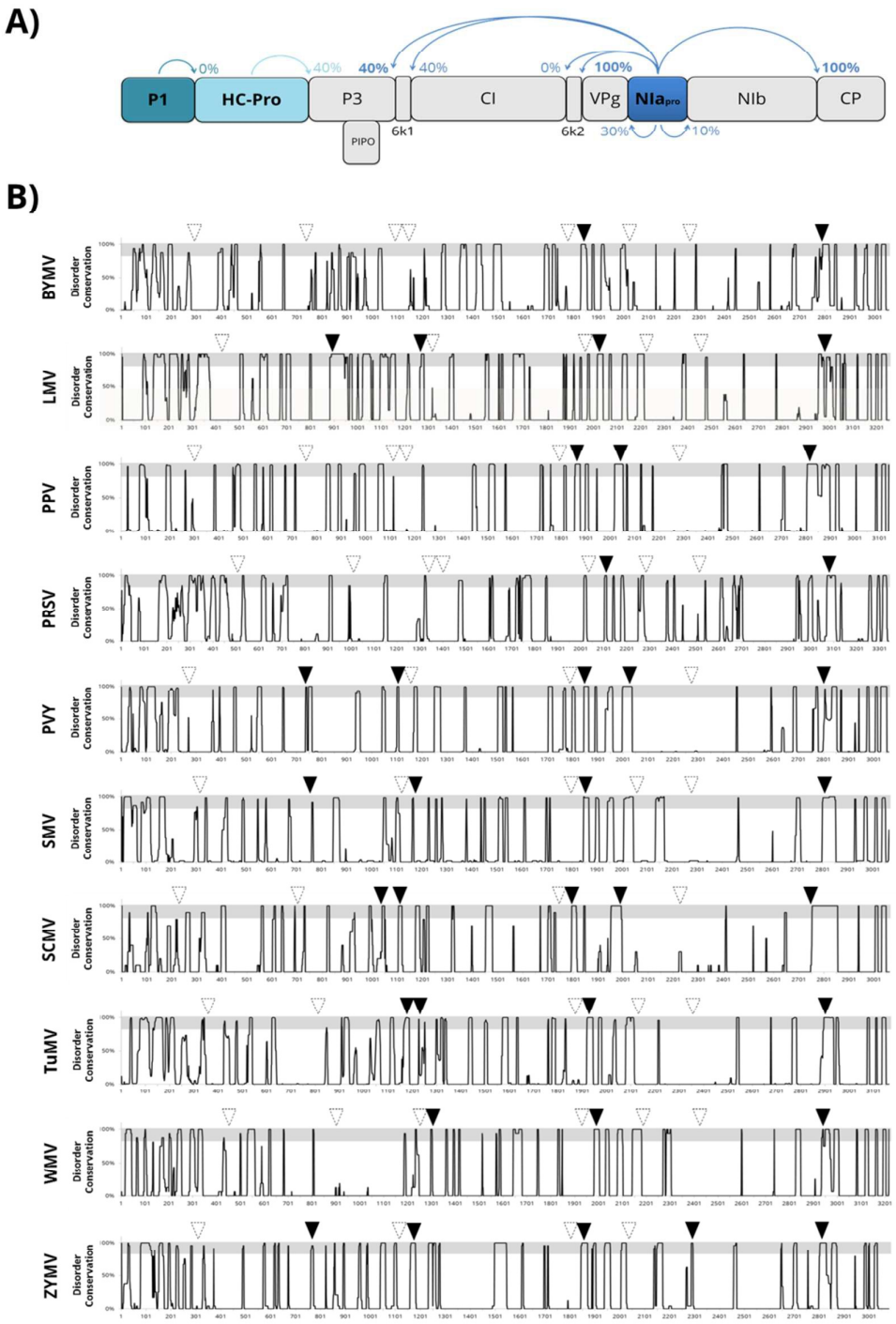


FIG 4. Disorder status of potyvirus polyproteins cleavage sites regions. A) Potyvirus polyprotein representation and cleavage sites. Values represent the percentage of disorder conservation at the level of each cleavage site for the ten potyvirus species. Cyan, light and dark blue arrows indicate sites processed by P1, HC-Pro and NIa-pro respectively. **B)** Polyprotein displayed disorder landscape. Disorder conservation scores obtained along the polyprotein for each species dataset (see Table S1 for sequences accessions). The grey shaded horizontal bar indicates the 80% cut-off used to define positions where disorder is conserved at the intra-species level. Triangles indicate cleavage sites that are in predicted disordered (dark ones), and ordered regions (white and dotted ones).

Charon et al. Figures

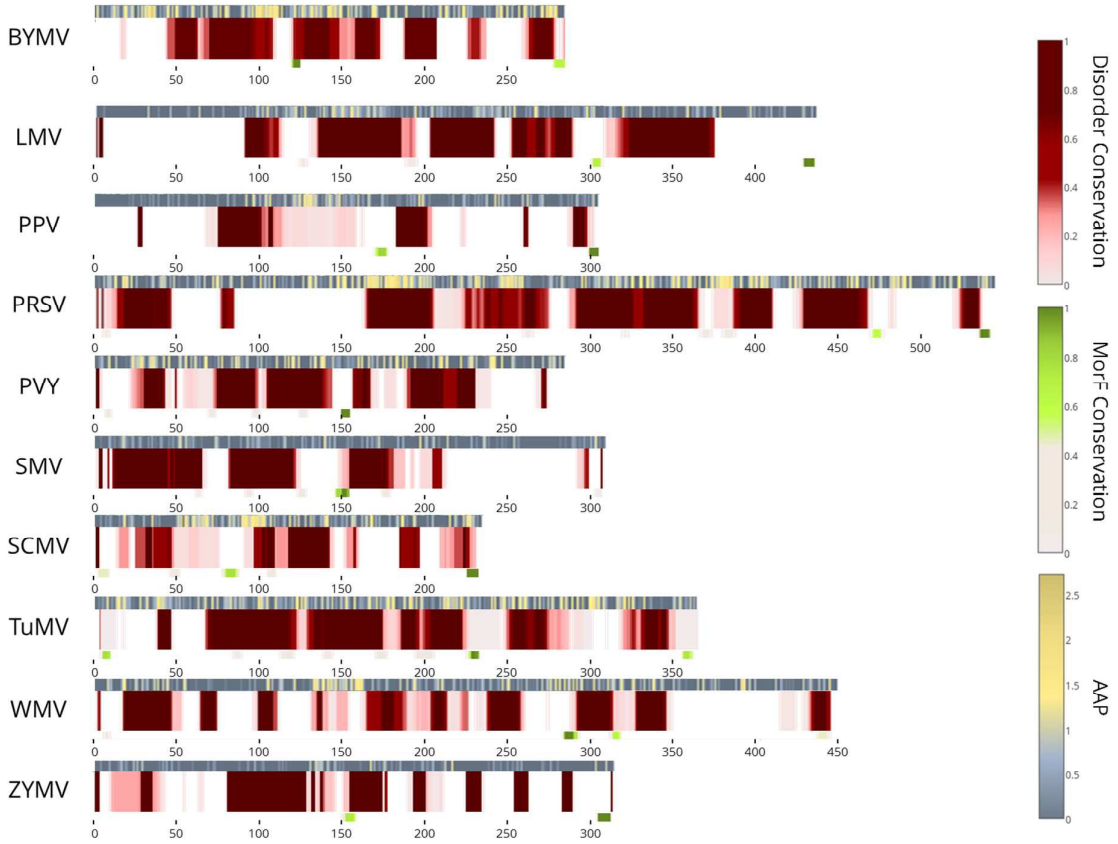


FIG 5. Conserved IDRs along P1 protein of ten potyvirus. White-to-red gradation bar represent degree of disorder conservation, from 0% (white) to 100% (dark red). White-to-green bar represent Molecular recognition features (MoRFs) conservation signal, from 0% (white) to 100% (dark green). Blue-to-yellow bar represents amino acid polymorphism (AAP). By definition, 0 to 1 represent highly conserved position, 1 to 2 is considered as moderately conserved and higher than 2 is considered as variable.