Order in Disorder as Observed by the "Hydrophobic Cluster Analysis" of Protein Sequences

Tristan Bitard-Feildel, Alexis Lamiable, Jean-Paul Mornon, Isabelle Callebaut

Order in disorder as observed by the "hydrophobic cluster analysis" of protein sequences

Tristan Bitard-Feildel[1,2], Alexis Lamiable[2], Jean-Paul Mornon[2], Isabelle Callebaut[2,*]

1. Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France
2. Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590, IRD, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, 75005 Paris, France

* To whom correspondence is to be sent:
isabelle.callebaut@upmc.fr

IMPMC, UMR7590

case 115, 4 place Jussieu, 75252 Paris Cedex 05, France

Abbreviations: HC: hydrophobic cluster, HCA: Hydrophobic Cluster Analysis, RSS: regular secondary structure, IDP: intrinsically disordered protein, IDR: intrinsically disordered region

Keywords: HCA, secondary structure, disorder, foldability, dark proteome

Total number of words (including references as well as figure and table legends) : 8335

Abstract

Hydrophobic Cluster Analysis (HCA) is an original approach for protein sequence analysis, which provides access to the foldable repertoire of the protein universe, including yet unannoted protein segments ("dark proteome"). Foldable segments correspond to ordered regions, as well as to intrinsically disordered regions (IDRs) undergoing disorder to order transitions. In this review, we illustrate how HCA can be used to give insight into this last category of foldable segments, with examples matching known 3D structures. After reviewing the HCA principles, we give examples of short foldable segments, which often contain small linear motifs, typically matching hydrophobic clusters. These segments become ordered upon contact with partners, with secondary structure preferences generally corresponding to those observed in the 3D structures within the complexes. Such small foldable segments are sometimes larger than the segments of known 3D structures, including flanking hydrophobic clusters that may be critical for interaction specificity or regulation, as well as intervening sequences allowing fuzziness. We also present cases of larger conditionally disordered domains, with lower density in hydrophobic clusters than well-folded globular domains or with exposed hydrophobic patches, which are stabilized by interaction with partners.

1. Introduction

Protein domains are structural and functional units which, through well-defined 3D structures, orchestrate various processes, from enzyme catalysis to signal transduction. Protein domains are evolutionary conserved at the sequence and structure level and several domain databases have been developed, providing statistical models which allow automatic protein annotation [1]. The use of protein domains in different contexts, a phenomenon called versatility or promiscuity, permits the molecular tinkering necessary for functional diversification and species evolution [2,3]. The presence or absence of domains in species can also be considered to track back molecular innovation over evolutionary time [4,5].

During the last two decades, it has become clear that the functional toolkit of proteins is not limited to well-structured domains, but also involves intrinsically disordered regions (IDRs), *i.e.* protein segments, and sometimes whole proteins (IDPs), which lack a stable, well-defined tertiary structure, at least in their native, unbound state [6,7]. IDRs are prevalent in eukaryotic sequences and occupy central positions in cellular interaction networks, fulfilling important regulatory, signaling, assembly and scaffolding roles. Recent works have highlighted their roles in newly discovered mechanisms, especially in the formation of membraneless organelles or biomolecular condensates by liquid-liquid phase separation, in which they provide multiple, weakly adhesive interacting elements [8-10]. Several definitions have been proposed depending on the functional or structural contexts in which IDPs/IDRs are considered and on the experimental techniques used to identify disorder. Different flavors of disorder can generally be distinguished. Molecular recognition involving IDRs is especially mediated by short motifs, constituting efficient, convergently evolvable solutions for interfaces [11-13] and conferring outstanding evolutionary plasticity to proteomes [14,15]. They enable low affinity, transient and conditional interactions, which can be easily modulated for instance but not exclusively, through post-translational modifications (PTMs) [16-18]. Short motifs are designated as linear motifs (LMs), eukaryotic linear motifs (ELMs) or short linear motifs (SLiMs) [19] and, more recently in the MobiDB database, as linear interacting peptides (LIPs) [20]. They often undergo disorder-to-order transitions when interacting with structured domains of partners [16] and in these cases, can also be described as preformed structural elements (PSEs) [21], molecular recognition elements (MoREs) or

molecular recognition features (MoRFs) [22-24], primary contact sites [25] or prestructured motifs (PreSMos) [26]. These preformed structural elements, likely representing binding competent states and displaying significant level of amino acid sequence conservation, are often embedded within fully disordered regions, a feature that was largely exploited for their detection from the sequence information [13,27].

Large-scale annotation and prediction of disorder have been the subject of many bioinformatics developments [20,28]. However, disorder predictors generally depend on the proxies that are used and may suffer from the scarcity of large benchmarking datasets, which are moreover heterogeneous. Also, they generally can not provide insights into disorder flavors which have not been yet described experimentally.

In this review, we focus on an approach, called Hydrophobic Cluster Analysis (HCA), which allows to delineate and get information about regions which are likely to be ordered (either in stable or conditional ways) as well as, by inference, disordered, from the only information of a single amino acid sequence. It provides a global view of the protein sequence texture, with insights into the structural features of foldable regions. After recalling the principles of HCA and related methodological approaches and databases, we provide the reader with guidelines to its use for delineating foldable regions, with special emphasis on cases of conditional order/disorder.

2. Hydrophobic Cluster Analysis and the delineation of foldable regions

Differences between order and disorder can be appreciated at the level of the amino acid sequence, as disordered regions are significantly depleted in order-promoting residues (W,C,F,I,Y,V,L,N) and enriched in disorder-promoting residues (A,R,G,Q,S,P,E,K) [29]. Order-promoting residues mostly include strong hydrophobic amino acids (V,I,L,M,Y,W,F), which mainly belong to regular secondary structures and participate to the densely packed cores of globular domains [30]. A very simple way to get information about “ordered” regions, from the only information of a single amino acid sequence, is to consider clusters made of these strong hydrophobic amino acids, as defined by Hydrophobic Cluster Analysis (HCA) [30,31]. HCA is based on a duplicated bidimensional (2D) representation of the protein sequence,

which highlights local proximities between amino acids [30,31] (**Figure 1**). Using a 2D net implies considering a connectivity distance (CD), which is the minimal number of positions required to interrupt the connectivity between amino acids. In the HCA representation, the sequence is written on a duplicated α-helical net (CD 4), in which strong hydrophobic amino acids (V, I, L, F, M, Y, W) are encircled and their adjacent contours joined, forming the so-called hydrophobic clusters (HCs) (**Figure 1**). As illustrated with the examples shown in **Figure 1** and assessed in a quantitative way from the analysis of experimental 3D structures datasets [32,33], these HCs mainly correspond to regular secondary structures (RSSs). The robustness of the chosen connectivity distance and hydrophobic alphabet in providing the best correspondence between HCs and RSSs has been assessed against sets of non-redundant, experimental 3D structures of globular domains [32,33]

Examination of an HCA plot, which can be drawn using the DrawHCA tool (**Table1**) thus gives, at a glance, information about the RSSs positions as well as their marked or more ambiguous preferences towards a particular state (see chapter 3 below). This information is gained from the only information of a single amino acid sequence, which is particularly useful for analyzing orphan sequences, *i.e.* sequences without any homologs. Moreover, a high density in HCs indicates the presence of foldable regions, corresponding to either soluble, globular domains or membrane domains, depending on their total content in hydrophobic amino acids and HC lengths [30]. Indeed, analysis of the SCOPe database (2.07) at 40 % redundancy indicates that globular domains *(classes a to e, 13293 proteins)* have on average 33.3 % of strong hydrophobic amino acids (standard deviation (SD) 3.7), with HC lengths up to 13-14 amino acids, whilst membrane domains, cell surface proteins and peptides *(class f, 271 proteins)* have a higher content in strong hydrophobic amino acids (mean 41.1 %, SD 9.2) and longer HCs. By contrast, regions lacking HCs or possessing only small and/or scarcely distributed HCs generally correspond to fully disordered sequences and/or flexible linkers. These features that can be deduced from HCA have been supported in a quantitative way by developing a tool, called SEG-HCA, allowing to automatically delineate regions with high density in HCs (foldable regions) [34,35]. The relevance of such approach has been supported by considering the coverage of domain and structure databases by the SEG-HCA predictions [34,35]. The vast majority of conserved domains are indeed well covered by SEG-HCA predictions (over up to 95 % of their lengths), the few ones

not being detected corresponding to domains with less hydrophobic amino acids and stabilized by metal ions or disulfide bridges. Applying SEG-HCA on whole proteomes allowed to comprehensively delineate foldable regions, corresponding to 85.5 % of the UniProt/SwissProt [36] (**Figure 2**, (A) blue and green sections, (B) blue and purple sections). This percentage has to be compared to the 61 % covered by Pfam (v31.0) domains [37], revealing that a large part (35.1 %) of the Pfam-unannotated sequences, also referred as the dark proteome [38], in fact corresponds to orphan foldable regions (**Figure 2-B**). Our studies, together with the work of Perdigao and colleagues [34,35,39] thus highlighted that the dark proteome has a limited amount of fully disordered proteins or segments (less than 4 %), contrary to some assumptions [40]. Orphan domains correspond either to "true" orphan sequences (*i.e.* sequences sharing no obvious similarity with any other sequence or domain (24.2 % and 63 % of UniProt/Swissprot orphan domains, respectively) or sequences sharing remote relationships with already known families of domains (12.7 % of Uniprot/Swissprot orphan domains), as systematically explored using sensitive bioinformatics tools [35]. Remote relationships can be detected considering 2D signatures defined by HCA, as illustrated by the identification of new families of domains starting from the analysis of orphan sequences (*e.g.* [41-46]). Bioinformatics tools have been developed to help such analysis [47]. Interestingly, the comprehensive analysis of whole proteomes has indicated that SEG-HCA predicted foldable regions can also be highlighted within the set of regions that are predicted as disordered using current disorder predictors, such as IUPRED [48] or MobiDB-lite [28] (green section in **Figure 2-A**). These regions generally correspond to protein segments undergoing disorder-to-order transitions [34] and correlate with ANCHOR predictions [49], which are based on pairwise energy estimation. They are generally short, foldable regions, having the ability to mediate transient interactions. These features are also found in PSEs, embedded within highly flexible carrier regions [13,21,50].

3. 2D structure content of foldable regions.

Hydrophobic clusters (HCs) can also give useful information about RSS type (α-helix or β-strand), based on the only information of a single amino acid sequence, without knowledge of any homologous sequence, thus making them particularly interesting to analyze orphan proteins. HCs can be described as non-overlapping binary patterns, defined as unique

combinations of hydrophobic (1) and non-hydrophobic (0) positions and separated from each other by at least four non-hydrophobic amino acids or a proline (**Figure 1**). They carry a more relevant information about RSSs than simple binary patterns, due to the consideration of this connectivity distance [51]. Each binary code defines a HC species, which can adapt a large variety of amino acid sequences. Secondary structures propensities and associated affinities (affinity corresponds to the RSS state for which the maximal propensities are observed) were calculated for the most frequent HC species considering experimental 3D structure databases. First limited to 294 frequent HC species [33], this database now contains a total of 476 frequent HC species (Supporting information, **Table S1**). 64.2 % of the total number of hydrophobic clusters found in UniProt/SwissProt fall into these 476 HC species, which cover 29.6 % of the sequence lengths (excluding from the calculations HC species 1 (a single hydrophobic amino acid, which is not preferentially associated with any RSS)). As illustrated with the two examples shown in **Figure 1,** some HC species have clear preference for α-helices (H) or β-strands (E) and have binary patterns typical of the periodicity observed in these RSSs. These binary pattern preferences have been supported in a comprehensive way over the whole set of HC species present in the experimental 3D structures of globular domains [52]. RSS prediction can be refined for HC with strong (E/H) but also moderate preferences (e/h) for any RSS by considering amino acid composition, as distinct amino acids profiles are observed for the two RSS states associated with each HC species [52].

4. A practical example of HCA-based delineation of foldable regions: the ENA/VASP protein

We first illustrate here the usefulness of HCA for predicting the foldable and disordered regions by expanding the example of enabled/vasodilator-stimulated phosphoprotein (Ena/VASP), a protein involved in actin assembly [53] (**Figure 3**). Five foldable regions (black boxes) are delineated on this sequence using the SEG-HCA program, four of which being experimentally characterized at the 3D structure level (grey boxes). The first and fifth foldable domains are large (>40 amino acids), match order predictions (as illustrated by the IUPRED [48] and consensus MobiDB-lite [28] predictions) and indeed correspond to stable 3D structures. The first globular domain (EVH1/WH1 domain) binds the linear motif FPPPP found in various VASP partners [54,55], whilst the fifth domain corresponds to a right-handed α-helical coiled-coil, allowing tetramerization [56]. The two other, smaller foldable regions

(third and fourth ones), included in disordered regions but matching ANCHOR predictions of disorder-to-order transitions [49,57], are typical examples of short linear motifs that fold upon binding to their partners. These two regions (making part of a larger EVH2 domain) are known as the globular and filamentous actin-binding sites (GAB and FAB) and are separated from the EVH1/WH1 domain by a proline-rich region, which binds profilin and the SH3 and WW domains of signaling and scaffolding proteins. Upon interaction with actin, GAB and FAB fold as α-helices, displaying structural similarities with the WH2 domain of WASP [58]. The two peptides, shown here on orange (Ena/VASP GAB motif) [53] and green (WH2 region of N-WASP, sharing structural similarities with the ENA/VASP FAB domain) [59], are shown within the complex with actin/profilin (grey). Of note is the overall good prediction of the limits of foldable regions when compared to experimental information. Moreover, good correspondences are globally observed between observed RSSs and predictions, particularly for clusters with strong affinities for RSSs (H and E), for which the binary pattern overwhelms the amino acid composition [52]. These predictions are based on the single amino acid sequence information (thus differing from current RSS predictors, based on amino acid profiles) and on the HC binary pattern information (**Table S1**). For those clusters that are more difficult to predict, the amino acid composition can help the prediction [52]. For instance, cluster with P-code 35 (h) in the EVH1 domain contains amino acids, such as V, I, T, C, S, which have preferences for extended structures. Considering some amino acids, such as A (α-helices) and T/C (β-strands), within the hydrophobic alphabet may also guide the analysis. This is for example the case of the GAB motif, including several alanine residues and made of the two HC basic units (called quarks, **Figure 1**) d and u, typical of helical conformation. Interestingly, the hydrophobic face of the GAB and FAB motifs, which undergo disorder-to-order transitions, complements the solvent-exposed hydrophobic patch of the binding partner. Too long clusters are not sufficiently represented in the 3D structure databases to allow relevant statistics (nd = not determined). However, some of these long clusters (see P-code 7269 in the EVH1 domain) can be split into two separate clusters (dotted red line), corresponding to two different RSSs. The structural behavior of other HCs can also be anticipated when they have clear horizontal shapes (thus HCs with Q-codes made of a majority of u and d), typical of α-helices or even coiled-coils (see the C-terminal tetramerization domain).

Thus, calculation of mean RSS propensities (mean of individual propensities for each amino acid) within hydrophobic cluster limits generally provides relevant predictions about the expected structural behavior of foldable regions, whenever these correspond to stable 3D structures or undergo disorder-to-order transitions.

5. Some examples of conditional disorder explored using HCA

In this section, we focus on specific cases of conditional disorder, illustrating how to apply the HCA approach in search of such protein segments. These examples have been selected by visual inspection of the experimental 3D structures of foldable motifs, extracted using SEG-HCA from the UniProt/SwissProt database, either being short (≤ 30 amino acids) or larger but having a lower hydrophobic content than stable, globular domains. A last example deals with complex cases of conditional disorder observed in protein globular-like domains with standard amino acid composition and specific 3D structure [60]. Note that a foldable segment, as detected by HCA, may correspond to an autonomous unit, folding in a stable or conditional way, but may also be part of a larger domain, being separated from the first one, at the sequence level, by large loops. Such a possibility can be inferred from a careful analysis of the sequence neighborhood of the foldable segment.

5.1. Short foldable segments

IDPs can be classified into separate categories, depending on the strength of the interaction they establish with their partners [19,61]. In case of relatively strong interaction, linear segments are multipartite, between 20 and 50 amino acids long, and consequently, interaction surface is relatively large (> 500 $Å^2$). Examples can be found of both intra- and intermolecular interactions. An example of a tight, intra-molecular interaction is illustrated here in **Figure 4-A** with the ever shorter telomeres 3 (Est3) protein, a regulatory OB-fold protein belonging to the yeast telomerase holoenzyme. The short foldable segment of Est3 is located in the N-terminus of the protein and make a spiral-shaped structure that caps the top of the OB barrel [62]. This region seems to be critical for telomerase function, as recently reported for its remote mammalian homolog TPP1 [63]. In a general way, IDRs appear to be a convenient tool used by auto-inhibited proteins for the fine-tuning of equilibrium between active and inactive states [64]. Some inter-molecular interactions mediated by foldable

segments also involve a relatively large surface of the partners, within large, multisubunit complexes, probably contributing to their stability or regulation. This is for instance the case of the N-terminal arm of methylmalonyl coA mutase α subunit, wrapping around the β-subunit (**Figure 4-B**). However, numerous inter-molecular interactions of foldable segments occur through limited surfaces, involving shorter sequence motifs (3-10 amino acids) and smaller surfaces (~500 $Å^2$) [19]. Several examples are illustrated on **Figures 4-C/4-D** (α-forming peptides) and **4-E/4-F** (β-forming peptides). In these examples, agreeing with previous observations [21], the predicted RSS preferences of the HCs involved in the interaction (as assessed by the affinity of the HC species) generally correspond to the RSSs observed in the complexes. This is particular true for species with strong RSS affinities (**Figures 4-D** and **4-E**, as well as **Figure 2** (FAB region)). In these examples, the hydrophobic amino acids of the HC complement the hydrophobic patch present at the partner surface. Worth noting is that the foldable segments boxed in **Figures 4-A, 4-C, 4-D, 4-E** and **4-F** are larger than the segments whose 3D structure has been solved (shaded in red), including more HCs than the one involved in the interaction. Examination of solvent accessible surfaces of the partner (**illustrated on Figure 4-C**) suggests that HC(s) flanking the interacting HC may dock into hydrophobic groove(s) present in close vicinity to the central binding site and may thereby reinforce or modulate the transient interaction. These SLIMs may thus be part of larger intrinsically disordered domains (IDDs), being multipartite [19]. There are also cases in which the affinity of the interacting HC does not correspond to the observed RSS, as illustrated with the Apollo (DCR1B) and SLX4 TRF2-binding motifs, which overlap the motif also present in Tin2 (**Figure 5**). In these examples, some hydrophobic amino acids of the interacting HC stay exposed to the solvent. The interacting HC is however also accompanied within the foldable segments by other HCs, which may interact together to form a small globular-like 3D structure. A similar situation is encountered for the Artemis (DCR1C) DNA ligase IV-binding peptide (aa 485 to 495 [65]) within the foldable segment encompassing aa 446 to 507 (data not shown).

Thus, considering the limits of foldable segments, as they can be predicted by visual inspection of HCA plots or through the SEG-HCA tool, may allow to clarify the structural boundaries of the SLIMs/IDDs and therefore to better understand the affinity and specificity of functional interactions, as well as of their fuzziness.

5.2. Larger, conditionally disordered domains

Disorder can also be observed for large foldable regions (*i.e.* of length > 50 amino acids) and can be classified in two categories. First, foldable segments which have less than 30-35 % of hydrophobic amino acids (percentage typical of globular domains, see before), presenting more sparsely distributed HCs, with large inter-HC regions. This is exemplified here with the N-terminal domain of coronavirus nucleocapsid N phosphoprotein, which provides a scaffold for viral RNA packaging. The domain is rich in basic amino acids, but has only 27 % of strong hydrophobic amino acids (of which several aromatic amino acids), thus less than the mean percentage of stable globular domains (**Figure 6**). Highly flexible loops disordered in the solution structure becomes ordered around a central β-sheet in the crystal lattice, a mechanism which may be critical for ribonuclecapsid assembly [66]. Second, there are also case of foldable segments which, despite a total content in hydrophobic amino acids typical of globular domains, seem unable to fold in a stable way, while homologs sharing similar sequences are stable and folded under similar conditions [60]. The expected 3D structure of the conditionally disordered domains, involving non-local sequences contacts, is then achieved by post-translational modifications or environmental perturbations, including specific binding partners. The gain of specific tertiary structures, and not only of secondary structures as observed for small linear interacting motifs, can thus be described as an extensive coupled folding and binding process. This is for example the case of the domain we detected in the C-terminus of the human AF9 and yeast TAF14 proteins, both members of the YEAST family, which shares significant similarity with the Extra-Terminal (ET) domain of BET (Bromo and Extra-Terminal) proteins [47], as illustrated by the conservation of hydrophobic clusters (shaded gray in **Figure 7**). Both families of proteins play key role in chromatin modification and transcription [67]. In the absence of the small interacting peptide of its partner AF4, the AF9 ET domain is indeed disordered [68], whilst the ET domain of BRD4 was found structured in isolation [69]. Hydrophobic residues of the AF4 linear interacting peptide, also undergoing coupled folding and binding and matching a small foldable region, complete the hydrophobic core of the AF9 ET domain by forming an inter-molecular three-stranded β-sheet (**Figure 7**). A similar mechanism is observed for the NSD3 peptide interacting with BDR3 and also matching a small foldable region. Noteworthy, the topology of the first hydrophobic cluster of the ET domain, with strong strand (E) affinity but corresponding to an α helix (α1)), is indicative of exposed hydrophobic amino acids and thus

of putative unstability and/or binding sites. Interestingly, several experimental 3D structures of the BRD3 and BRD4 ET domains were recently solved in complex with the small interacting peptides from different partners, again matching well small foldable regions (bottom panel of Figure 7). These structures highlight a versatile common binding pocket, able to accommodate peptides in different conformations [70-73] (**Figure 7**). The most common effector recognition mode is through antiparallel β-sheet formation (involving one or two β-strands of the partner). However, in the BRD4/JMJD6 complex, the JMJD6 linear peptide retains a helical conformation similar to that observed in the full JMJD6 protein (helix α6) and interacts with the BRD4 ET three-helix bundle [71]. Of note is that in contrast to other cases, the JMJD6 small interacting peptide is not embedded within flexible linkers, but is included into a well-folded domain. Interaction with BRD4 ET domain would thus require significant conformational rearrangement of JMJD6, likely occurring upon binding to single-stranded RNA [71]. The binding platform provided by ET domains is probably critical for the recruitment of several chromatin remodeling complexes and transcription regulators to promoters and enhancers. The functional advantages of the relative lack of stability and flexibility of such small, folded domains might be linked to the modulation of binding rates and affinities for the different partners. Interestingly, examination of the AF9 and BRD3/4 HCA plots (**Figure 7**) indicate two possible, yet uncharacterized small foldable segments, upstream of their respective ET domains, with strong propensities for α-helical secondary structures (black stars). These peptides could possibly form intramolecular interactions with the ET domains, allowing to stabilize them in absence of their interacting partners.

6. Conclusions

HCA is an *ab-initio* approach that can be used in addition to current disorder prediction tools, as described in some reviews [6,74-76]. **Table 1** provides a list of tools integrating the HCA concepts for order/disorder prediction and visualization. Even though most of the works using HCA have been focused on well-folded domains, with several ones dealing with the identification of new families of domains starting from the analysis of orphan sequences (*e.g.* [41-46]), several studies have more particularly explored disorder [77-81], with special emphasis on proteins from viruses [82-84] or from parasites [85] and plant proteins involved in various responses [86-88]. These applications underscore the interest of the HCA approach

especially for analyzing orphan proteins, common in proteomes with amino acid compositional bias. This bias generally leads to spurious, non-relevant protein sequence matches when using standard tools for similarity search, while leaving relevant ones undetected.

In this context, identifying short linear motifs that fold upon binding is a challenging task due to the fact that these are often embedded within highly variable, disordered sequences. HCA-based analyses are of interest as they only need the information of a single amino sequence and do not suffer from the statistical uncertainties associated with sequence similarity searches. Once the foldable segments have been identified, they can be then further explored for potential similarities, searched at the level of the amino acid sequence or at the level of hydrophobic clusters, which are much more conserved than the sequence itself. Hydrophobic clusters indeed constitute structural signatures as the hydrophobic character of about one half of the hydrophobic amino acids composing them is conserved in homologous sequences of globular domains, in which they participate in the protein cores[89]. Such signatures can thus be used to identify specific signals within a highly noised background, thus at very low levels of sequence identity, as illustrated for instance by the HCA-based detection of hidden transcription factors associated with RNA pol II in Apicomplexan proteomes [85]. This proven strategy in the case of globular domains is also of interest for short linear motifs that fold upon binding, also known as MORFs, as their interfaces are characterized by a high hydrophobicity, complementing hydrophobic patches on the surface of the partner proteins [24].

Short linear motifs bind their target proteins with sufficient strength to establish a functional interaction and adopt a defined structure upon binding. However, if the bonds between the linear peptides and their targets are sufficient to ensure binding, they are too few to explain the high degree of specificity observed *in vivo*. It is thus the biological context which determines interaction specificity. This information is, to a great extent, contained in the residues lying outside the short linear motifs. Moreover, these flanking residues play an important role in the conformational heterogeneity maintained upon interaction, a general behavior which is described as fuzziness [90] and which has been analyzed in the vicinity of linear peptides [91,92]. Context residues are described as allowing specificity, in particular by

preventing cross reactions (negative selection) while more flexibility is allowed. We suggest here, based on several examples, that the foldable segments delineated using the HCA/SEG-HCA approach may allow to clarify the structurally relevant limits of interacting segments, including the flanking hydrophobic cluster(s), beyond the immediate vicinity of the hydrophobic cluster of the central linear motif. These additional short hydrophobic motifs may thus be used in combination in order to enhance specificity or binding strength, a multipartite binding mechanism which has already been documented [61]. Discontinuous binding motifs may then be separated by parts of the segments which remain disordered, allowing fuzziness [90]. A comprehensive survey of linear interacting peptides reported in databases will allow to further understand the importance of the hydrophobic cluster neighborhood. A detailed analysis of the enrichment of linear interacting peptides in specific HC species will also provide useful information for their prediction at the level of whole proteomes.

Acknowledgments

This work was supported by grants from the Agence Nationale de la Recherche Scientifique (ANR-14-CE10-0021, ANR-14-CE14-0028 and ANR-17-CE12-0016) and from the Institut National du Cancer (2014-1-PL BIO-09 and 2016-PL BIO-11).

References

[1] A. Scaiewicz, M. Levitt, *Curr Opin Genet Dev* **2015**, *35*, 50.

[2] J. Jin, X. Xie, C. Chen, J. Park, C. Stark, D. James, M. Olhovsky, R. Linding, Y. Mao, T. Pawson, *Sci Signal* **2009**, *2(98)*, ra76.

[3] E. Bornberg-Bauer, M. M. Albà, *Curr Opin Struct Biol* **2013**, *23*, 459.

[4] X. C. Zhang, Z. Wang, X. Zhang, M. H. Le, J. Sun, D. Xu, J. Cheng, G. Stacey, *BMC Evol Biol* **2012**, *12*, 6.

[5] K. Forslund, A. Henricson, V. Hollich, E. L. Sonnhammer, *Mol Biol Evol* **2008**, *25*, 254.

[6] J. Habchi, P. Tompa, S. Longhi, V. N. Uversky, *Chem Rev* **2014**, *114*, 6561.

[7] 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. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, M. M. Babu, *Chem Rev* **2014**, *114*, 6589.

[8] S. F. Banani, H. O. Lee, A. A. Hyman, M. K. Rosen, *Nat Rev Mol Cell Biol* **2017**, *18*, 285.

[9] S. Boeynaems, S. Alberti, N. L. Fawzi, T. Mittag, M. Polymenidou, F. Rousseau, J. Schymkowitz, J. Shorter, B. Wolozin, L. Van Den Bosch, P. Tompa, M. Fuxreiter, *Trends Cell Biol* **2018**, *28*, 420.

[10] A. L. Darling, Y. Liu, C. J. Oldfield, V. N. Uversky, *Proteomics* **2018**, *18*, e1700193.

[11] V. Neduva, R. B. Russell, *FEBS Lett* **2005**, *579*, 3342.

[12] N. E. Davey, G. Trave, T. J. Gibson, *Trends Biochem Sci* **2011**, *36*, 159.

[13] N. Davey, K. van Roey, R. J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel, T. J. Gibson, *Mol Biosyst* **2012**, *8*, 268.

[14] P. Tompa, N. Davey, T. J. Gibson, M. M. Babu, *Mol Cell* **2014**, *55*, 161.

[15] R. Pancsa, P. Tompa, *Trends Biochem Sci* **2016**, *41*, 896.

[16] P. E. Wright, H. J. Dyson, *Curr Opin Struct Biol* **2009**, *19*, 31.

[17] K. Van Roey, T. J. Gibson, N. E. Davey, *Curr. Opin. Struct. Biol* **2012**, *22*, 378.

[18] A. L. Darling, V. N. Uversky, *Front Genet* **2018**, *9*, 158.

[19] K. van Roey, B. Uyar, R. J. Weatheritt, H. Dinkel, M. Seiler, A. Budd, G. T. J, N. E. Davey, *Chem Rev* **2014**, *114*, 6733.

[20] D. Piovesan, F. Tabaro, L. Paladin, M. Necci, I. Micetic, C. Camilloni, N. Davey, Z. Dosztányi, B. Mészáros, A. M. Monzon, G. Parisi, E. Schad, P. Sormanni, P. Tompa, M. Vendruscolo, W. F. Vranken, S. C. E. Tosatto, *Nucleic Acids Res* **2018**, *34*, 445.

[21] M. Fuxreiter, I. Simon, P. Friedrich, P. Tompa, *J Mol Biol* **2004**, *338*, 1015.

[22] C. J. Oldfield, Y. Cheng, M. S. Cortese, P. Romero, V. N. Uversky, A. K. Dunker, *Biochemistry* **2005**, *44*, 12454.

[23] A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker, V. N. Uversky, *J Mol Biol* **2006**, *362*, 1043.

[24] V. Vacic, C. J. Oldfield, A. Mohan, P. Radivojac, M. S. Cortese, V. N. Uversky, A. K. J. Dunker, *Proteome Res* **2007**, *5*, 2351.

[25] V. Csizmok, M. Bokor, P. Bánki, E. Klement, K. Medzihradszky, P. Friedrich, K. Tompa, P. Tompa, *Biochemistry* **2005**, *44*, 3955.

[26] S. H. Lee, D. H. Kim, J. J. Han, E. J. Cha, J. E. Lim, Y. J. Cho, C. Lee, K. H. Han, *Curr Protein Pept Sci* **2012**, *13*, 34.

[27] F. Meng, V. N. Uversky, L. Kurgan, *Cell Mol Life Sci* **2017**, *74*, 3069.

[28] M. Necci, D. Piovesan, Z. Dosztányi, S. C. E. Tosatto, *Bioinformatics* **2017**, *33*, 1402.

[29] 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. Nissen, R. Reeves, C. Kang, C. R. Kissinger, E. C. Garner, Z. Obradovic, *J Mol Graph Mod* **2011**, *19*, 26.

[30] I. Callebaut, G. Labesse, P. Durand, A. Poupon, L. Canard, J. Chomilier, B. Henrissat, J. P. Mornon, *Cell. Mol. Life Sci.* **1997**, *53*, 621.

[31] C. Gaboriaud, V. Bissery, T. Benchetrit, J. P. Mornon, *FEBS Lett.* **1987**, *224*, 149.

[32] S. Woodcock, J. P. Mornon, B. Henrissat, *Protein Eng* **1992**, *5*, 629.

[33] R. Eudes, K. Le Tuan, J. Delettre, J. P. Mornon, I. Callebaut, *BMC Struct Biol* **2007**, *7*, 2.

[34] G. Faure, I. Callebaut, *PLoS Comput Biol* **2013**, *9*, e1003280.

[35] T. Bitard-Feildel, I. Callebaut, *Sci Rep* **2017**, *7*, 41425.

[36] T. Bitard-Feildel, I. Callebaut, *bioRxiv* **2018**, n/a.

[37] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, A. Bateman, *Nucleic Acids Res* **2016**, *44*, D279.

[38] M. Levitt, *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 11079.

[39] N. Perdigao, J. Heinrich, C. Stolte, K. S. Sabir, M. J. Buckley, B. Tabor, B. Signal, B. S. Gloss, C. J. Hammang, B. Rost, A. Schafferhans, S. I. O'Donoghue, *Proc Natl Acad Sci U S A* **2015**, *112*, 15898.

[40] A. Bhowmick, D. H. Brookes, S. R. Yost, H. J. Dyson, J. D. Forman-Kay, D. Gunter, M. Head-Gordon, G. L. Hura, V. S. Pande, D. E. Wemmer, P. E. Wright, T. Head-Gordon, *J Am Chem Soc* **2016**, *138*, 9730.

[41] I. Callebaut, J. C. Courvalin, J. P. Mornon, *FEBS Lett* **1999**, *446*, 189.

[42] I. Callebaut, J. de Gunzburg, B. Goud, J. Mornon, *Trends Biochem Sci* **2001**, *26*, 79.

[43] I. Callebaut, J. Mornon, *FEBS Lett* **1997**, *400*, 25.

[44] I. Callebaut, J. Mornon, *Bioinformatics* **2005**, *21*, 699.

[45] I. Callebaut, J. Mornon, *Bioinformatics* **2010**, *26*, 1140.

[46] I. Callebaut, D. Moshous, J. P. Mornon, J. P. de Villartay, *Nucleic Acids Res* **2002**, *30*, 3592.

[47] G. Faure, I. Callebaut, *Bioinformatics* **2013**, *29*, 1726.

[48] Z. Dosztányi, V. Csizmók, P. Tompa, I. Simon, *J Mol Biol* **2005**, *347,* 827.

[49] Z. Dosztányi, B. Mészáros, I. Simon, *Bioinformatics* **2009**, *25*, 2745.

[50] M. Fuxreiter, P. Tompa, I. Simon, *Bioinformatics* **2007**, *23*, 950.

[51] J. Hennetin, K. Le Tuan, L. Canard, Colloc'h, N, J. P. Mornon, I. Callebaut, *Proteins* **2003**, *51*, 236.

[52] J. Rebehmed, F. Quintus, J. Mornon, I. Callebaut, *Proteins* **2016**, *84*, 624.

[53] F. Ferron, G. Rebowski, S. H. Lee, R. Dominguez, *EMBO J* **2007**, *26*, 4597.

[54] K. E. Prehoda, D. J. Lee, W. A. Lim, *Cell* **1999**, *97*, 471.

[55] L. Ball, R. Kühne, B. Hoffmann, A. Häfner, P. Schmieder, R. Volkmer-Engert, M. Hof, M. Wahl, J. Schneider-Mergener, r. U. Walte, H. Oschkinat, T. Jarchau, *EMBO J* **2000**, *19*, 4903.

[56] K. Kühnel, T. Jarchau, E. Wolf, I. Schlichting, U. Walter, A. Wittinghofer, S. V. Strelkov, *Proc Natl Acad Sci U S A* **2004**, *101*, 17027.

[57] B. Mészáros, I. Simon, Z. Dosztányi, *PLoS Comput Biol* **2009**, *5*, e1000376.

[58] D. Chereau, F. Kerff, P. Graceffa, Z. Grabarek, K. Langsetmo, R. Dominguez, *Proc Natl Acad Sci USA* **2005**, *102*, 16644.

[59] J. F. Gaucher, C. Maugé, D. Didry, B. Guichard, L. Renault, M. F. Carlier, *J Biol Chem* **2012**, *287*, 34646.

[60] A. C. Hausrath, R. L. Kingston, *Cell Mol Life Sci* **2016**, *74*, 3149.

[61] P. Tompa, *Curr Opin Struct Biol* **2011**, *21*, 419.

[62] T. Rao, J. W. Lubin, G. S. Armstrong, T. M. Tucey, V. Lundblad, D. S. Wuttke, *Proc Natl Acad Sci U S A* **2014**, *111*, 214.

[63] S. Grill, V. M. Tesmer, J. Nandakumar, *Cell Rep* **2018**, *22*, 1132.

[64] T. Trudeau, R. Nassar, A. Cumberworth, E. T. Wong, G. Woollard, J. Gsponer, *Structure* **2013**, *21*, 332.

[65] C. Charlier, G. Bouvignies, P. Pelupessy, A. Walrant, R. Marquant, M. Kozlov, P. De Ioannes, N. Bolik-Coulon, S. Sagan, P. Cortes, A. Aggarwal, L. Carlier, F. Ferrage, *J Am Chem Soc* **2017**, *139*, 12219.

[66] K. S. Saikatendu, J. S. Joseph, V. Subramanian, B. W. Neuman, M. J. Buchmeier, R. C. Stevens, P. Kuhn, *J Virol* **2007**, *81*, 3913.

[67] J. Schulze, A. Wang, M. Kobor, *Epigenetics* **2010**, *5*, 573.

[68] B. Leach, A. Kuntimaddi, C. Schmidt, T. Cierpicki, S. Johnson, J. Bushweller, *Structure* **2013**, *21*, 176.

[69] Y. Lin, T. Umehara, M. Inoue, K. Saito, T. Kigawa, M. Jang, K. Ozato, S. Yokoyama, B. Padmanabhan, P. Güntert, *Protein Sci* **2008**, *17*, 2174.

[70] D. C. Wai, T. N. Szyszka, A. E. Campbell, C. Kwong, L. E. Wilkinson-White, A. P. G. Silva, J. K. K. Low, A. H. Kwan, R. Gamsjaeger, J. N. Chalmers, W. M. Patrick, B. Lu, C. R. Vakoc, G. A. Blobel, J. P. Mackay, *J Biol Chem* **2018**, *293*, 7160.

[71] T. Konuma, D. Y. Yu, C. Zhao, Y. Ju, R. Sharma, C. Ren, *Sci Rep* **2017**, *7*, 16272.

[72] B. L. Crowe, R. C. Larue, C. Yuan, S. Hess, M. Kvaratskhelia, M. P. Foster, *Proc Natl Acad Sci U S A* **2016**, *113*, 2086.

[73] Q. Zhang, L. Zeng, C. Shen, Y. Ju, T. Konuma, C. Zhao, C. R. Vakoc, M.-M. Zhou, *Structure* **2016**, *24*, 1201.

[74] F. Ferron, S. Longhi, B. Canard, D. Karlin, *Proteins* **2006**, *65*, 1.

[75] V. Receveur-Bréchot, J.-M. Bourhis, V. N. Uversky, B. Canard, S. Longhi, *Proteins* **2006**, *62*, 24.

[76] B. He, K. Wang, Y. Liu, B. Xue, V. N. Uversky, A. K. Dunker, *Cell Res* **2009**, *19*, 929.

[77] D. S. Libich, M. Schwalbe, S. Kate, H. Venugopal, J. K. Claridge, P. J. B. Edwards, D. Dutta, S. M. Pascal, *FEBS J* **2009**, *276*, 3710.

[78] B. Xue, A. K. Dunker, V. N. Uversky, *J Biomol Struct Dynam* **2012**, *29*, 843.

[79] P. Mendoza-Espinosa, D. Montalvan-Sorrosa, V. Garcıa-Gonzalez, A. Moreno, R. Castillo, J. Mas-Oliva, *Mol Cell Biochem* **2014**, *393*, 99.

[80] E. M. Fernandez, M. D. Díaz-Ceso, M. Vilar, *PLoS ONE* **2015**, *10*, e0117206.

[81] C. Myrum, A. Baumann, H. J. Bustad, M. Innselset Flydal, V. Mariaule, S. Alvira , G. Cuellar , J. Haavik, J. Soulé, J. M. Valpuesta , J. A. Marquez, A. Martinez, C. R. Bramham, *Biochem J* **2015**, *468*, 145.

[82] J. Habchi, L. Mamelli, H. Darbon, S. Longhi, *PLOS One* **2010**, *5*, e11684.

[83] B. Da Costa, S. Duquerroy, B. Tarusa, B. Delmas, *Virus Research* **2011**, *158,* 251.

[84] A. Benarouche, J. Habchi, A. Cagna, O. Maniti, A. Girard-Egrot, J. Cavalier, S. Longhi, F. Carriere,

Biophys J **2017**, *113*, 2723.

[85] I. Callebaut, K. Prat, E. Meurice, J. P. Mornon, S. Tomavo, *BMC Genomics* **2005** *6*, 100.

[86] L. N. Rahman, V. V. Bamm , J. A. M. Voyer, G. S. T. Smith, L. Chen, M. W. Yaish, B. A. Moffatt , J. R. Dutcher, G. Harauz, *Amino Acids* **2011**, *40*, 1485.

[87] X. Sun, D. R. Greenwood, M. D. Templeton, D. S. Libich, T. K. McGhie, B. Xue, M. Yoon, W. Cui, C. A. Kirk, W. T. Jones, V. N. Uversky, E. H. A. Rikkerink, *FEBS J* **2014**, *281*, 3955.

[88] K. Hamdi, E. Salladini, D. P. O'Brien, S. Brier, A. Chenal, I. Yacoubi, S. Longhi, *Sci Rep* **2017**, *7*, 15544.

[89] A. Poupon, J. P. Mornon, *Proteins* **1998**, *33*, 329.

[90] P. Tompa, F. M, *Trends Biochem Sci* **2008**, *33*, 2.

[91] A. Stein, P. Aloy, *PLoS One* **2008**, *3*, e2524.

[92] C. Chica, F. Diella, T. J. Gibson, *PLoS One* **2009**, *4*, e6052.

[93] P. Lieutaud, B. Canard, S. Longhi, *BMC Genomics* **2008**, *9*, S25.

[94] F. Ferron, C. Rancurel, S. Longhi, C. Cambillau, B. Henrissat, B. Canard, *J Gen Virol* **2005**, *86*, 743.

[95] D. Piovesan, I. Walsh, G. Minervini, S. Tosatto, *Bioinformatics* **2017**, *33*, 1889.

[96] K. Coeytaux, A. Poupon, *Bioinformatics* **2005**, *21*, 1891.

[97] C.-T. Su, C.-Y. Chen, C.-M. Hsu, *Nucleic Acids Res* **2007**, *35*, W465.

[98] F. Mancia, P. R. Evans, *Structure* **1998**, *6*, 711.

[99] I. Deshpande, A. Seeber, K. Shimada, J. J. Keusch, H. Gut, S. M. Gasser, *Mol Cell* **2017**, *68*, 431.

[100] B. D. Darimont, R. L. Wagner, J. W. Apriletti, M. R. Stallcup, P. J. Kushner, J. D. Baxter, R. J. Fletterick, K. R. Yamamoto, *Genes Dev* **1998**, *12*, 3343.

[101] C. y. Chang, J. D. Norris, H. Grøn, L. A. Paige, P. T. Hamilton, D. J. Kenan, D. Fowlkes, D. P. McDonnell, *Mol Cell Biol*. **1999**, *19*, 8226.

[102] V. Buzón, L. R. Carbó, S. B. Estruch, R. J. Fletterick, E. Estébanez-Perpiñá, *Mol Cell Endocrinol* **2012**, *348*, 394.

[103] K. Jehle, L. Cato, A. Neeb, C. Muhle-Goll, N. Jung, E. W. Smith, V. Buzon, L. R. Carbó, E. Estébanez-Perpiñá, K. Schmitz, L. Fruk, B. Luy, Y. Chen, M. B. Cox, Bräse, S, M. Brown, A. C. Cato, *J Biol Chem* **2014**, *289*, 8839.

[104] N. Yan, J.-W. Wu, J. Chai, W. Li, Y. Shi, *Nat Struct Mol Biol* **2004**, *11*.

[105] E. M. Romes, A. Tripathy, K. C. Slep, *J Biol Chem* **2012**, *287*, 15862.

[106] Y. Chen, Y. Yang, M. van Overbeek, J. R. Donigian, P. Baciu, T. de Lange, M. Lei, *Science* **2008**, *319*, 1092.

[107] B. Wan, J. Yin, K. Horvath, J. Sarkar, Y. Chen, J. Wu, K. Wan, J. Lu, P. Gu, E. Y. Yu, N. F. Lue, S. Chang, Y. Liu, M. Lei, *Cell Rep* **2013**, *4*, 861.

HCA-based tools		
DrawHCA	HCA plots	*http://obsornite.impmc.upmc.fr/hca/hca-form.html* *http://mobyle.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py#forms::HCA*
SEG-HCA	Foldable region delineation	[34,36]
TREMOLO-HCA	Remote homology detection using 2D signatures and domain architecture	[47]
MeDor (MEtaServer of DisORder)	Disorder prediction	[93]
VaZyMolO	Definition of modularity in viral proteins	[94]
FELLs (Fast Estimator of Latent Local Structure)	Visualization (SEG-HCA foldable segments)	[95]
Other prediction/analysis tools considering information extracted from hydrophobic clusters		[96,97]

Table 1:
Tools integrating the HCA concepts for order/disorder prediction and visualization

Legends to figures

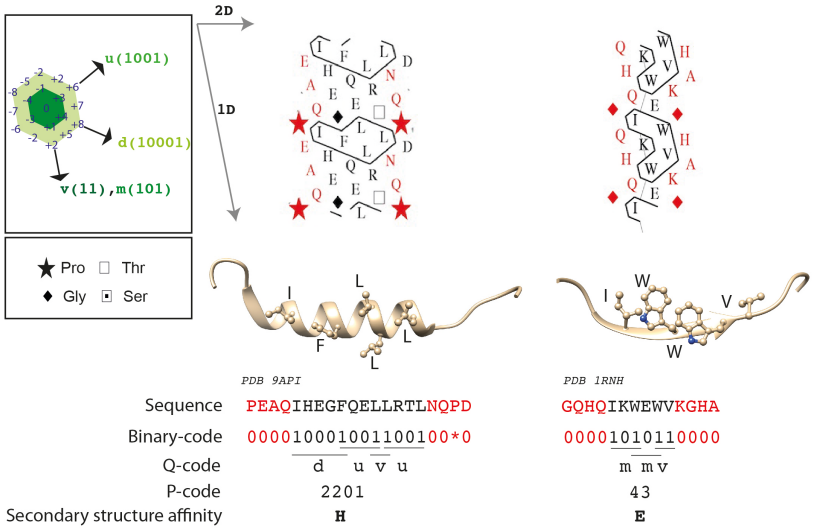

Figure 1: Principles of Hydrophobic Cluster Analysis. *The amino acid sequence is written on a duplicated α-helical net, in which the seven strong hydrophobic amino acids (V,I,L,M,F,Y,W) are contoured, forming hydrophobic clusters, which mainly correspond to regular secondary structures (RSSs). Hydrophobic clusters are separated from each other by at least four non-hydrophobic amino acids or a proline (amino acids depicted in red). The 2D net and neighborhood are detailed at left, together with the four symbols used for amino acids with particular structural behavior. At right are shown two examples of hydrophobic cluster (HC) species (each species being defined by a unique binary pattern) with strong affinities for α-helices (H) and β-strands (E), respectively, and the corresponding binary codes, Quark (Q)-codes and Peitsch (P)-codes. Quarks correspond to the four basic units (v (vertical, 11), m (mosaic, 101), u (up 1001) and d (down, 10001)), from which any hydrophobic cluster can be built. The three axes corresponding to these quarks are shown at left on the 2D net. P-codes correspond to the sums of powers of 2, indexed according to the position of each number of the binary code (the last position of the hydrophobic cluster corresponding to 0).*

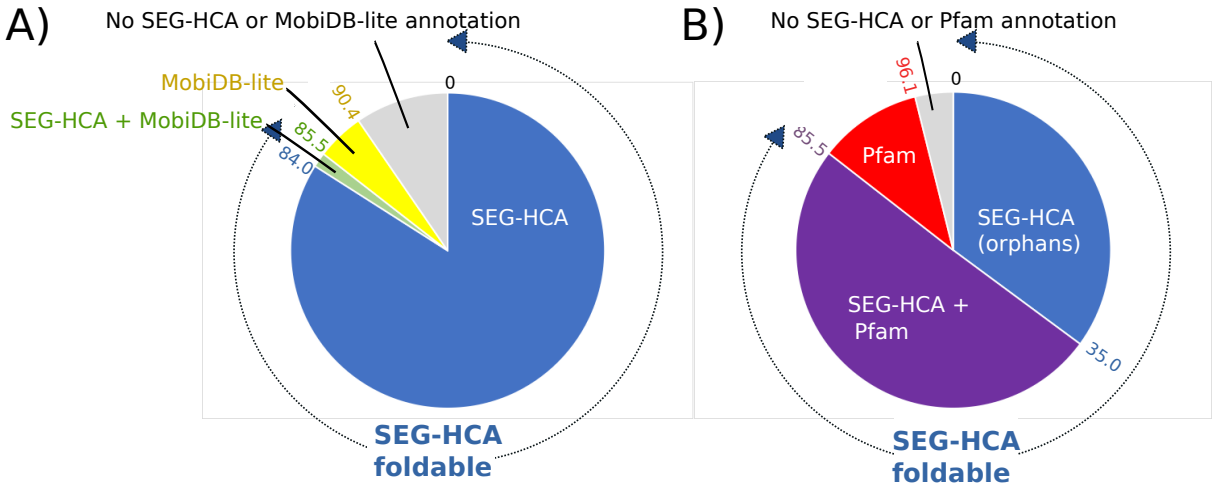

Figure 2: Amino acid coverage of the UniProt/SwissProt database by SEG-HCA foldable regions. *These predictions are compared to (A) consensus disorder predictions, as made by MobiDB-lite* [28] *and (B) domain database annotations (Pfam v31.0)* [37].

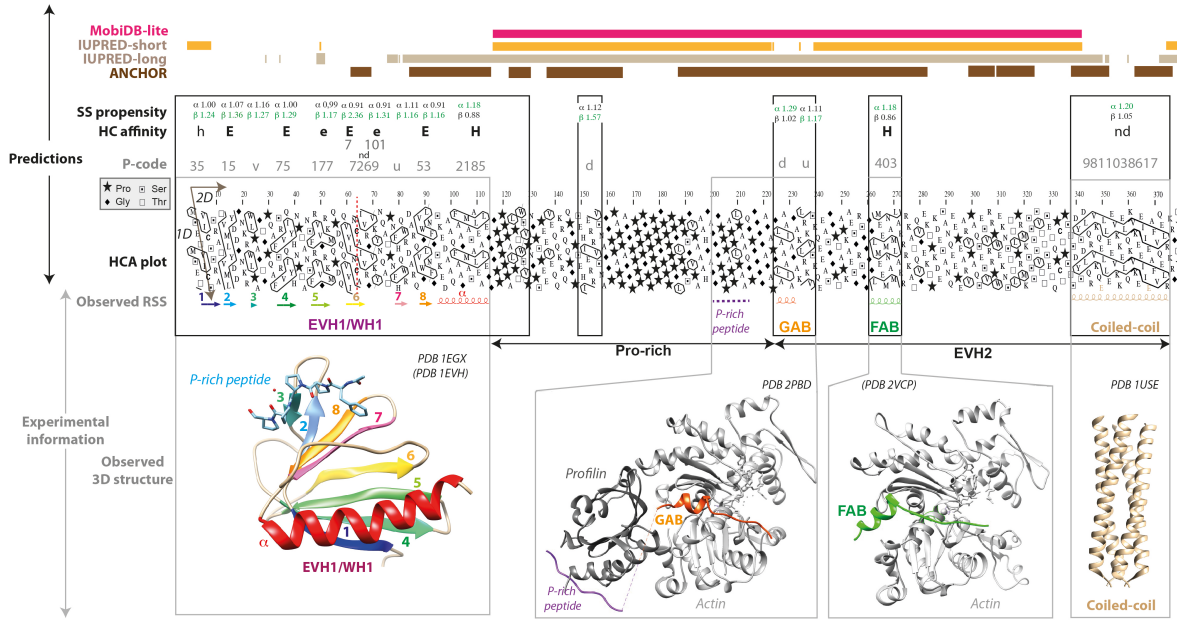

Figure 3: Delineation of order and disorder in the human enable/vasodilator-stimulated phosphoprotein (Ena/VASP – UniProt P50552). *The foldable regions, as predicted using SEG-HCA are boxed (black) on the HCA plot. Additional information is reported about the corresponding experimental data (observed 3D structures and corresponding RSSs) (grey boxes, with PDB identifiers indicated) and order/disorder predictions (upper part). Colored bars: predictions of disorder reported by MobiDB-lite (consensus) [20], as well as by IUPRED [48] and by ANCHOR (disorder-to-order transitions) [49,57]. Peitsch (P-)codes and HC affinities for RSS are indicated (E/e: strand, H/h: helix, with upper/lower cases corresponding to strong and weak affinities, respectively), except for the four basic units (called "quarks", see Figure 1), displaying per se no clear secondary structure affinities. No statistics (nd=not determined) are available for too long clusters, which can however sometimes be split into more informative, shorter clusters (dotted red bar). RSS propensities focused on the HC limits (mean of the individual propensities of each amino acids for the different RSS) generally provide relevant predictions about the expected structural behavior (highest propensities are shown in green).*

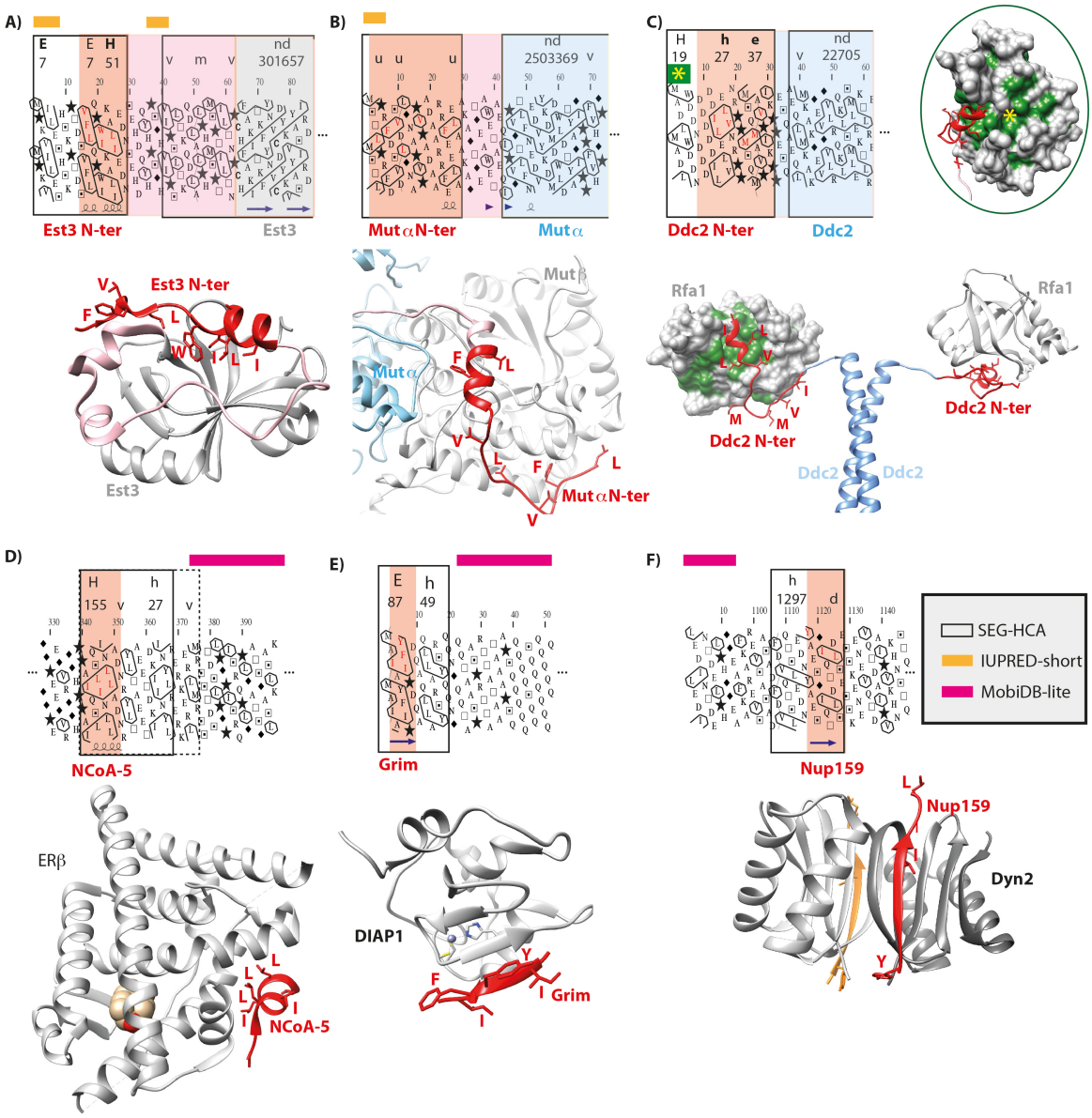

Figure 4: Short foldable segments

On the HCA plots, the positions of foldable segments delineated using SEG-HCA are boxed, whereas those of the corresponding interacting peptide 3D structures found within small foldable segment are shaded in red. These interacting peptides are depicted in red on the ribbon representation of the 3D structure complexes, with the hydrophobic amino acids depicted in atomic details. The interacting partner is depicted in grey. Observed RSS and predictions are indicated below of or up to the HCA plots, respectively.

A-B. Long peptides. A) Intra-molecular interaction. *The N-terminal region of the Est3 telomerase subunit, forming together with the C-terminal region, a cap covering a 5-stranded β-barrel (UniProt Q03096, PDB 2M9V [62]).* ***B) Inter-molecular interaction****. The N-terminal arm of the methylmalonyl coA mutase α subunit, wrapping around the β-subunit (UniProt P11653, PDB 3REQ [98]).*

C-F. Short linear motifs. C) The Replication Protein A (RPA)-binding domain of S. cerevisiae Ddc2 (UniProt Q6CUV9, ATRIP in human) in complex with the N-terminal OB fold of the RPA's largest subunit (S. cerevisiae Rfa1, RPA70 in human) ([99], PDB 5OMC). *The N-terminal*

region of Ddc2 serves as a RPA-binding domain allowing the recruitment of the Mec1-Dcd2 complex (ATR-ATRIP in human), a key DNA-damage-sensing kinase, to DNA damage sites [99]. The additional hydrophobic cluster, upstream the interacting hydrophobic cluster, may bind to the hydrophobic extension of the binding groove, depicted at right on the solvent accessible surface (yellow star). ***D) The LXXLL motif (NR box) of the rat nuclear receptor coactivator (NCoA-5, UniProt Q9HCD5) in complex with estrogen receptor beta ERβ (PDB 2J7X).*** *The α-helicoidal LXXLL motif fits into a groove of the ERβ ligand-activated hormone binding domain (AF-2 pocket). Flanking sequences of LXXLL NR boxes have been shown to be involved in the modulation of the affinity and/or selectivity of interaction [100,101]. It is also possible here that the hydrophobic cluster downstream the NR box plays a role in the selectivity of the interaction or its regulation. This is supported by the fact that another druggable BF-3 pocket, conserved among nuclear receptors, has also been identified in the proximity of the AF-2 pocket [102], which has been shown to be targeted by NR-binding motifs [103].* ***E) The N-terminal IAP-binding motif of the D. melanogaster cell death protein Grim (UniProt Q24570) in complex with the first BIR (baculoviral IAP repeat) domain of Diap1, a member of the inhibitor of apoptosis family ([104], PDB 1SE0).*** *The pro-death protein Reaper, Hif and Grim (RHG) induce apoptosis by antagonizing DIAP1 function, by relieving the DIAP1-mediated inhibition of the effector caspase DrICE.* ***F) A peptide from the nuclear pore Nup159 (uniProt P40477), in complex with the core β-sandwich of the nucleoporin Dyn2, forming a homodimer ([105], PDB 4DS1).***

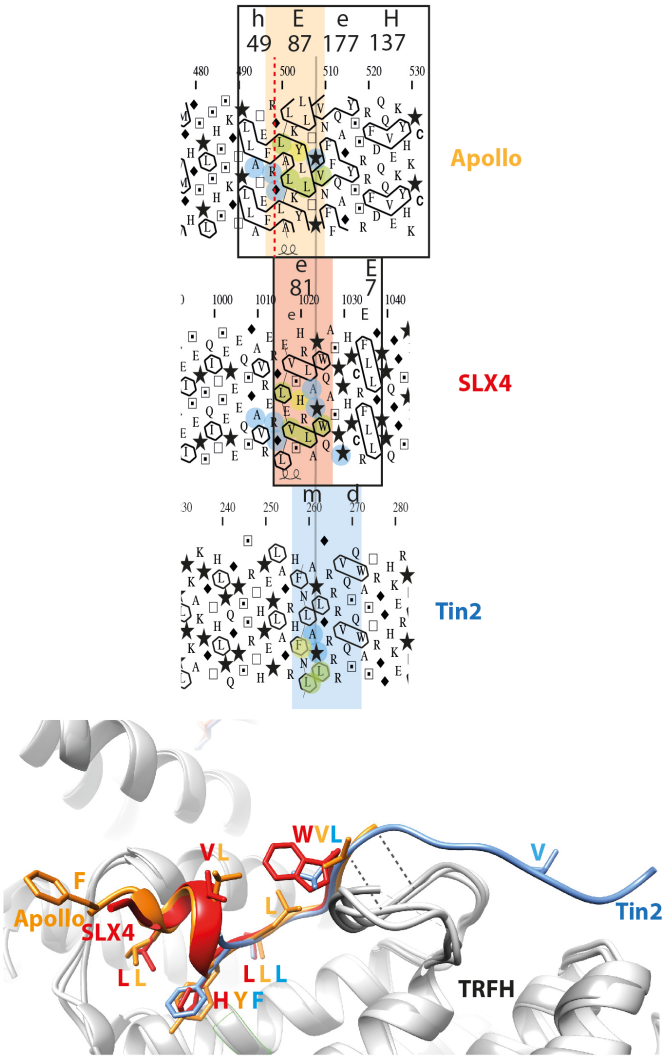

Figure 5. TRFH-binding motif (TBM). *The TBM of human SLX4 (UniProt Q8IY92) in complex with TRF2 (PDB 4M7C, [106]), compared to the TBM of Apollo (UniProt Q9H816) and of TIN2 (UniProt Q9BSI4) in complex with TRF2 and TRF1, respectively ([107], PDB 3BUA and 3BU8). The telomere restriction fragment homology (TRFH) domains of shelterin proteins TRF1 and TRF2 are the principal mediators that recruit several non-shelterin accessory proteins to telomeres. Of these are the SLX4 and Apollo nucleases, which share a short peptide with a common signature sequence YxLxP (red and orange), folding as an α-helix (sequence identities/similarities are shaded). The TRFH TIN2-interaction site is adjacent (blue), but distinct from the SLX4-Apollo binding site, with TIN2 binding in an extended conformation. Of note is that the first part of the TIN2 peptide perfectly superimposes with the end of the SLX4-Apollo peptides (see the corresponding sequence identities/similarities), suggesting that the segment C-terminal of the interacting peptide of SLX4 and/or Apollo might bind in an extended conformation in this adjacent site. This hypothesis is further supported by the fact that hydrophobic clusters with strand affinities are found downstream of the interacting peptide in the SLX4 and Apollo foldable segments delineated by SEG-HCA (red and grey boxes, respectively). The Tin2 peptide (shaded blue) was not detected as a putative foldable segment.*

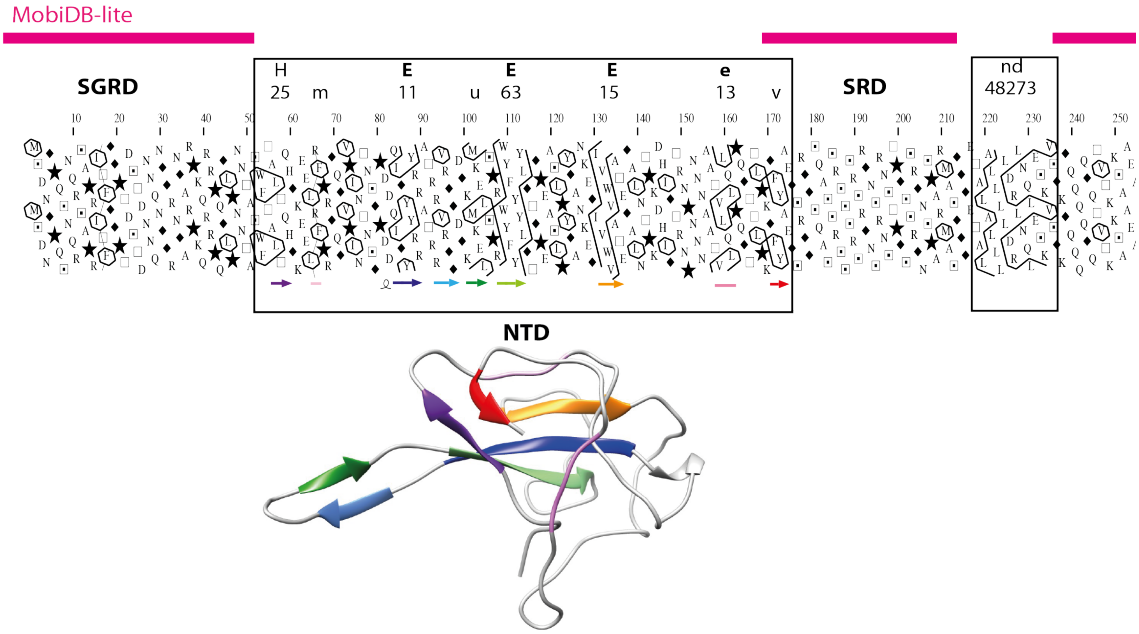

Figure 6: Large, disordered foldable segments, with a low density in hydrophobic clusters. *HCA plot of nucleoprotein of human SARS coronavirus (UniProt P59595) and crystal structure of the N-terminal domain (NTD, PDB 2OFZ). SGRD: Serine-glycine-arginine rich Domain, SRD: Serine rich Domain.*

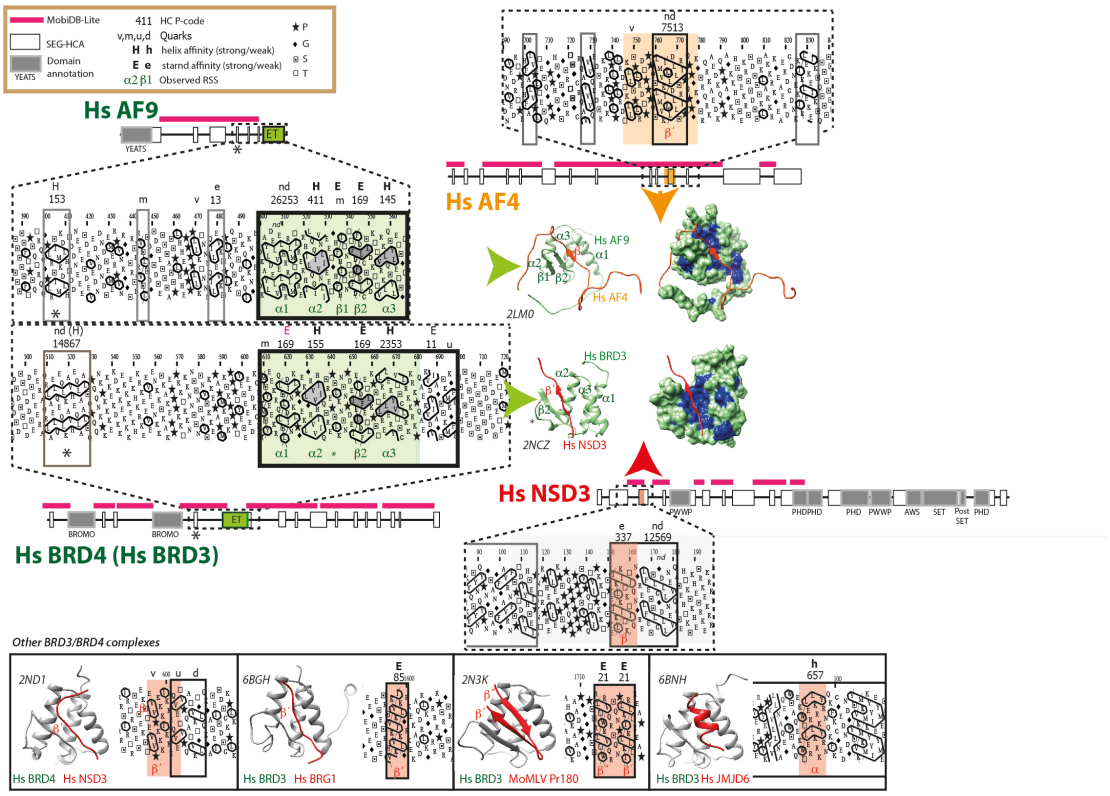

Figure 7: Large, conditionally disordered foldable domains, with standard density in hydrophobic clusters. *HCA plots of ET domains from the YEAST (top – human AF9) and BRDT (bottom - human BRD4) families, and their small interacting peptides in different protein partners (at right: human AF4 and human NSD3, as well as at bottom: a second peptide in human NSD3, human BRG1, MoMLV Pr180 and human JMJD6). Foldable regions, as predicted by SEG-HCA, are boxed, and the limits of observed 3D structures is shaded in green (ET domain) and in orange/red (small interacting peptides). These sequences are placed within the context of the whole protein architectures, for which are also reported PROSITE domain annotations, as well as MobiDB-Lite disorder annotations. Ribbon representations of the 3D structures are displayed, together with solvent accessible surface representations of the ET domain, illustrating the hydrophobic patch (blue) recognized by the interacting peptides. UniProt: Hs AF9: P42568 ,Hs BRD4: O60885, Hs NSD3: Q9BZ95, MoMLV (Moloney Murine Leukemia Virus) Pr180 (gag-Pro-Pol polyprotein): Q8UN00, Hs AF4: P51825, Hs BRG1: P51532, Hs JMJD6: Q6NYC1.*