A topology-based investigation of protein interaction sites using Hydrophobic Cluster Analysis

Alexis Lamiable, Tristan Bitard-Feildel, Joseph Rebehmed, Flavien Quintus, Françoise Schoentgen, Jean-Paul Mornon, Isabelle Callebaut

▶ **To cite this version:**

Alexis Lamiable, Tristan Bitard-Feildel, Joseph Rebehmed, Flavien Quintus, Françoise Schoentgen, et al.. A topology-based investigation of protein interaction sites using Hydrophobic Cluster Analysis. Biochimie, 2019, 167, pp.68-80. ⟨10.1016/j.biochi.2019.09.009⟩. ⟨hal-02323013⟩

HAL Id: hal-02323013

https://hal.science/hal-02323013v1

Submitted on 17 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

A topology-based investigation of protein interaction sites using Hydrophobic Cluster Analysis

Alexis Lamiable[1,a], Tristan Bitard-Feildel[1,b], Joseph Rebehmed[1,2], Flavien Quintus[1], Françoise Schoentgen[1], Jean-Paul Mornon[1], Isabelle Callebaut[1]*

1. Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, 75005 Paris, France
2. Lebanese American University, Department of Computer Science and Mathematics, Beirut, Lebanon

a Present address : Institut de Biologie de l'ENS, 75005 Paris

b Present address : Direction Générale de l'Armement - Maîtrise de l'Information (DGA-MI). 35170 Bruz

* Corresponding author:
Isabelle Callebaut
Isabelle.Callebaut@upmc.fr
IMPMC, UMR7590 CNRS Sorbonne Université
Case 115, 4 place Jussieu
75252 Paris Cedex 05, France

ABSTRACT

Hydrophobic clusters, as defined by Hydrophobic Cluster Analysis (HCA), are conditioned binary patterns, made of hydrophobic and non-hydrophobic positions, whose limits fit well those of regular secondary structures. They were proved to be useful for predicting secondary structures in proteins from the only information of a single amino acid sequence and have permitted to assess, in a comprehensive way, the leading role of binary patterns in RSS preference towards a particular state. Here, we considered the available experimental 3D structures of protein globular domains to enlarge our previously reported hydrophobic cluster database (HCDB), almost doubling the number of hydrophobic cluster species (each species being defined by a unique binary pattern) that represent the most frequent structural bricks encountered within protein globular domains. We then used this updated HCDB to show that the hydrophobic amino acids of discordant clusters, *i.e.* those less abundant clusters for which the observed secondary structure is in disagreement with the binary pattern preference of the species to which they belong, are more exposed to solvent and are more involved in protein interfaces than the hydrophobic amino acids of concordant clusters. As amino acid composition differs between concordant/discordant clusters, considering binary patterns may be used to gain novel insights into key features of protein globular domain cores and surfaces. It can also provide useful information on possible conformational plasticity, including disorder to order transitions.

Key words: amino acid sequence, hydrophobicity, secondary structures, globular domains, intrinsically disordered proteins, conformational plasticity.

1. INTRODUCTION

Stable, globular domains of proteins invariably fold into densely packed cores by burying **hydrophobic side chains** and exposing polar side chains to solvent [1, 2]. Hydrophobic amino acids are mostly found into **regular secondary structures** (RSSs, α-helices and β-strands) [3], in which the buried polar carbonyl and amide groups of the polypeptide backbone can form hydrogen bonds, thereby neutralizing their polarity [4]. The sequence periodicity of polar/non polar amino acids has been recognized, through specific examples, as playing a critical role in the choice between α-helices and β-strands [5-9] and has been largely considered for *de novo* protein design [4, 10]. We have previously supported these **binary pattern preferences** in an exhaustive way through the consideration of constrained binary patterns (also called hydrophobic clusters, as defined by Hydrophobic Cluster Analysis (HCA) [3, 11]), which particularly well match the limits of RSSs [12, 13]. This property is due to the consideration of a connectivity distance, linked to a two-dimensional (2D) representation of the protein sequence and corresponding to the minimal number of non-hydrophobic amino acids separating two hydrophobic clusters from each other (**Figure 1**). We indeed demonstrated that some **hydrophobic cluster species** (each species being defined by a unique binary pattern, independent from the amino acid composition) display high propensities for either the α- or β-states [14], and that the 180 most common species with high propensities for one of these two states tend to have different **binary patterns**, characterized by periodicities in hydrophobic amino acids typical of α-helices (hydrophobic amino acids every 3 or 4 residues) or β-strands (hydrophobic amino acids every 2 residues), respectively [15]. These species with high propensities for the α- and β-states (two of which being represented in **Figure 1**) represent two thirds of the most frequently observed species in globular domains, which together gather 81 % of the total number of hydrophobic clusters [14]. By considering simple topology rules for these hydrophobic clusters within the context of globular domains (**Figure 1**), an agreement (or concordance) between the expected (from the binary pattern) and observed RSS thus implies that hydrophobic amino acids are shielded from solvent while participating in the hydrophobic core. In contrast, a conflict (disagreement or discordance) means that some hydrophobic amino acids are exposed to solvent and prone to interact with partners.

Here, we performed a comprehensive analysis of 3D structures in order to evaluate whether hydrophobic amino acids belonging to discordant hydrophobic clusters are indeed more exposed to solvent and more involved in protein interfaces than those belonging to concordant hydrophobic clusters. To that aim, we first updated our hydrophobic cluster database (HCDB), leading to a more than two-fold increase of the number of representative hydrophobic clusters relative to HCDB v1 [14]. We then supported our assumption in a quantitative way and illustrated this general observation with several specific examples. Combining this binary pattern information together with that of the amino acid composition of hydrophobic clusters, which differs between concordance and discordance states, sheds light on yet unexplored topological features of protein interaction sites. It also opens perspectives for highlighting possible conformational switches, including disorder-to-order transitions.

2. MATERIAL AND METHODS

2.1. STRUCTURE DATABASE

Starting from the SCOPe (Structural Classification of Proteins-extended [16]) database 2.06 (95% redundancy), we performed the same filtering procedure as described in [14], which consisted in keeping only structures belonging to the first five SCOPe classes (all α, all β, α/β, α+β, multidomains), and only X-ray structures with a resolution lower than 2.5 Å. We also excluded 3D structures with more than 10% missing residues or structures that only contained C_α coordinates. The resulting database contains 18602 3D structures. Non-standard amino acids have been converted into their respective unmodified forms (Note that 92.8% of the conversions were from seleno-methionine (MSE) to methionine (M), the full details are given in **Supplementary Data 1**).

2.2. HYDROPHOBIC CLUSTER AND HYDROPHOBIC CLUSTER SPECIES DEFINITIONS

Hydrophobic clusters (HCs) were defined as previously, as a succession of strong hydrophobic amino acids (V I L F M Y W) separated from each other by breakers. Breakers are composed of at least four consecutive non-hydrophobic amino acids (connectivity distance) or a proline (**Figure 1**). Note that the robustness of the hydrophobic alphabet and of the two-dimensional representation (based on an α-helical net) has been previously assessed, providing the best correspondence between HC and

regular secondary structure (RSS) positions [3, 12]. HCs with different amino acid sequences are grouped into HC species provided that they share a unique binary pattern (*e.g.* the two sequences (amino acid one-letter code) shown in **Figure 2** : p**L**A**I**E**L**S**L**p and p**Y**T**I**S**L**N**I**dcsr belong to the same HC species (same binary pattern **1**0**1**0**1**0**1**, the amino acids corresponding to the breakers are indicated with lower cases).

2.3. SECONDARY STRUCTURE ASSIGNMENT

We used two different methods for assigning secondary structures from atomic coordinates: DSSP [17] and STRIDE [18]. We reduced the 8 or 7 states to only strand (E), helix (H) and coil (C) states, as described previously [14]. As the results are similar for the two assignments methods, we only present data obtained using the DSSP assignments. HC are then associated with secondary structures following the **OPS rule** (*One Position is Sufficient*), as previously reported [14, 15]. Accordingly, a HC is considered associated with a helix or a strand if at least one of its amino acids is assigned to H or E, respectively (**Figure 2A**). If a cluster contains both H and E amino acids, then it is associated with a state called M, for *multiple*. This assignment rule was preferred over a rule considering all the positions of the clusters, as also made in [14]. Indeed, it allows to overcome the limitations of the non-strict correspondence between HC and RSS limits, as well as to consider more than one RSS (helix or strand) covering a single HC (*multiple* state). The reliability of such an assignment was assessed by calculating the coverage of the clusters by the assigned secondary structure.

2.4. PROPENSITIES OF HC SPECIES TOWARDS SECONDARY STRUCTURE STATES

2.4.1. General rule

We computed propensities of a species for a secondary structure state (C, E, H or M) in a way similar to that followed for calculating amino acid propensities, as previously done [14]. Hence, the propensity for a HC in HC species X to be assigned to secondary structure state R is $\boldsymbol{P_{X,R} = (|X_R|/|R|)/(|X|/N)}$, where $|X_R|$ is the number of HCs in HC species X assigned to state R, $|X|$ the number of representative HCs in the HC species, $|R|$ the total number of HCs associated to state R in the database, and N the total number of HCs of all classes.

However, this formula does not take into account the varying lengths of the HC species binary codes. Since our protocol assigns the M structure state to HCs containing both an amino assigned to H and another one assigned to E, the longer a HC is, the more likely it is to be assigned to class M, as shown in **Supplementary Data 2**.

To take this bias into account, we reformulated the propensity formula in order to make it more independent on the HC length. We performed logistic regressions on the distributions depicted in **Supplementary Data 2**, in order to estimate the number of clusters of length *L* assigned to structure state *R*. The propensity formula then becomes $\boldsymbol{P_{X,R} = |X_R|/(|X|/N_{L,R})}$, where *L* is the length of HC *X*. $\boldsymbol{N_{L,R}}$ is computed by the logistic function, with the following parameters reported , computed from the database:

$\boldsymbol{N_{L,R}} = 1/\ (1+\exp(-(a_R \times L+b_R)\times scaling_R)$

Class	a	b	scaling
H	0.537	0.072	0.511
E	1.667	-0.142	0.595
C	2.558	-0.607	0.281
M	-4.772	0.406	0.485

2.4.2. Affinity assignment

To determine the affinity of a HC species *X* for a secondary structure, thus the secondary structure with the highest propensity, we use the following procedure (**Figures 2B-C**): for each secondary structure state *R* (H, E, C or M), we compute $\boldsymbol{P_{X,R}}$. We then compute the sum *T* of all the propensities, and the difference *D* between the highest and the second propensities. HC species *X* is defined as having an affinity for the class of highest propensity. If $D/T \geq 0.3$, then it is a high affinity, denoted by an uppercase letter. Otherwise it is a soft affinity, written as a lowercase letter.

2.4.3. Status

The notion of concordance or discordance of HCs in regard to their secondary structure affinity was introduced in [15]. A HC is considered concordant (abbreviated "**A**" for *Agreeing with the expected behavior*) with the affinity *R* of its HC species if at least 80% of its amino acids are assigned to *R.* It is considered discordant (abbreviated "**D**" for

Disagreeing with the expected behavior) if less than 20% are assigned to *R*, and intermediate (abbreviated "**I**") otherwise (**Figure 2D**).

We also defined concordance and discordance at the level of the HC species (**Figure 2E**). If R_X is the structural affinity of HC species *X*, as defined above, A_X and D_X the number of concordant (A) and discordant (D) HCs belonging to *X*, respectively (intermediate ones are ignored), then the HC species is considered strong (s) if $A_X \geq D_X$, weak (w) otherwise. Moreover, if $|A_X\text{-}D_X|/(A_X+D_X) \geq 20\%$, the HC species is considered highly strong or weak (written in uppercase). Hence, a HC species can be either S(trong), s(trong), W(eak) or w(eak).

3. RESULTS

3.1. HYDROPHOBIC CLUSTER DATABASE (HCDB V2)

A first version of the hydrophobic cluster database (**HCDB v1**) was published ten years ago [14]. This database was established by considering hydrophobic clusters (hereinafter "HC") extracted from **SCOP v1.69** (95% redundancy), from which 7015 protein chains of globular domains were kept after filtering. Sequence redundancy was treated at the level of each HC species, each HC species being defined by a unique binary code (or its decimal translation, called the Peitsch- or P-code) [14] (**Figure 1**), as well as by a unique Quark- or Q-code (decomposition of the HC into its basic units, called the Quarks, which are designated according to the direction followed in the two-dimensional representation of the sequence: 11 (*v* for vertical), 101 (*m* for mosaic), 1001 (*u* for up) and 10001 (*d* for down)) [15] (**Figure 1**). This treatment was made to avoid statistical bias, while preserving a sufficient number of clusters for each considered species. HCDB v1 contained the **294 most frequent hydrophobic clusters species** encountered in globular domains**,** with at least 30 occurrences and representing 81 % of the total number of clusters, at the time of its conception. HC species were characterized by the number of observed clusters (occurrence), the frequencies in the three secondary structures (SS) states (H-helix, E-strand and C-coil) and associated propensities, as well as their resulting SS affinity towards a particular SS state (thus corresponding to the maximal propensity). Among the 294 HC species included in HCDB v1, **97 and 83** appeared predominantly associated with the two regular SS (α-helix and

β-strand structures, respectively), after their propensities being calculated for the different states. HC species can be further divided in HC species with high affinities for a given secondary structure (upper cases H and E) and species with low affinities (lower cases h and e).

Since, the number of experimental 3D structures has significantly increased, allowing to get statistics on less populated HC species (especially of longer lengths) that were absent in the 2007 database. We have here followed the same protocol for updating the HC database (**HCDB v2**), starting from **SCOPe v2.06 95%** (28010 structures) [16]. After filtering (see **Material and methods**), **291542 HCs (22213 HC species)** were extracted from the amino acid sequences associated with **18602 protein chains**. The sequence redundancy was treated as previously, by considering less redundant versions of the SCOPe 95% database and evaluating by HC species the HC occurrence increase as a function of redundancy. If the difference in HC occurrence between two consecutive redundancy levels is greater than 10%, reflecting the presence of very similar sequences, only the structures included in the lowest redundancy level are considered for calculation. As reported previously [16], selecting HC from high redundancy banks allows to extend sets of informative species to higher lengths, as HC features are remarkably insensitive to redundancy. **476 HC species** are now present in HCDB v2 with more than 30 members (excluding the "1" HC species, which is more frequently encountered in the coil state, χ^2 statistical tests << 0.001), representing 82.7 % of the total number of HCs. As previously done [14], the OPS rule ("one position is sufficient") was adopted for assigning the HC species frequencies relative to each SS. Following this rule, a HC is associated with a α-helix or a β-strand if one or more contiguous amino acids of the HC are assigned H or E, respectively (**Figure 2A**). This allows overcoming weakened signals due to non-exact correspondence between HCs and RSSs. However, despite this simplified assignment, a good coverage of HCs by RSSs is observed, as in HCDB v1 (mean of cov values = 72 %, each cov value corresponding to the average number, calculated by HC species, of the HC residues assigned H or E within the HC limit, mean for all HC species: 72%). HCs associated with two RSS are considered apart and are designated as multiple (M). The complete HCDB v2 can be found at **http://osbornite.impmc.upmc.fr/HCADB/dictionary**, while an excerpt with some representative HC species is presented in **Figure 3A**.

Frequencies of the different HC species in the different SS states were used to calculate the corresponding propensities and associated affinities (**Figures 2B and 3A**). An improvement has been introduced in order to limit the bias towards the M state for the larger HC lengths (see **Material and Methods**). Among the 476 HC species now present in HCDB v2, **225 and 219** appeared predominantly associated with α-helices and β-strands, respectively (**Figure 4A**). HC species can be further divided in species with high affinities for a given SS (upper cases H (**108 species**) and E (**84 species**)) and species with low affinities (lower cases h (**117 species**) and e (**135 species**)) (**Figures 2C and 3A**). Relative to HCDB v1, the increase of the number of HC species with at least 30 members is noticeable from length 8 onwards and concerns species with affinities for the α-helix state but also for the β-strand state (**Figure 4B**). A significant increase in β-strand HC species (from 41 to 95 HC species with strong and low affinities) is indeed observed up to length 12, whereas the number of α-helix HC species grows from length 9 to 13. As in the HCDB v1 and somewhat unexpectedly, no sufficiently populated HC species is found with a length ≥ 14. This can however be explained by the fact that long clusters are not favored due to their excessive length relative to the maximal diameter of globular domains. This comprehensive analysis thus highlights the usual length limits of hydrophobic clusters, correlated to the maximal length of RSSs [17] and to the maximal allowed size of foldable domains [19]. Further increase of the number of 3D structures in PDB should not dramatically modify the HCDB v2 content, all the more that the number of new folds and new single-domain families is saturating [20, 21].

32 HC are assigned as "multiple" in HCDB v2, of which 10 were also present and assigned in this state in HCDB v1. The largest part of HCDB v1 "multiple" HC (102 out of 112) are now assigned in HCDB v2 to the α-helix and β-strand states. This represents an improvement relative to HCDB v1, in which these multiple affinities were overrepresented. Of note is that 82 of these 102 "old multiple" HC species evolve towards low affinities for secondary structures (lower cases h and e), indicating that these HCs don't have straight structural behaviors. Affinities of 129 HC species are strictly conserved between HCDB v1 to v2, of which 78 correspond to strong affinities for the α-helix (H) and β-strand (E) states. Thus, 78 out of the 88 E/H HC species of HCDB v1 have strictly conserved their affinities, the ten others shifting towards low affinities. 29 other HC species shift from low (HCBD v1) to strong (HCBD v2) affinities

towards either the α-helix (H) or β-strand (E) states. Finally, 13 HC species shift between the α-helix and β-strand states, but this concern only low affinities (h, e) in HCDB v1 and v2. Only one case of a shift between the α-helix and β-strand states is observed with a high affinity for the α-helix state (H) in HCDB v2 (P-201, 11001001, HCDB v1 e). The HCDB v2 H affinity appears coherent given the inspected assignments of individual HCs within this HC species and its Q-code (*vuu*), mixing Quarks typical of the β- (*v*) and α- (*u*) states [15]. Thus, although some differences can be observed due to the data increase and treatment of multiple HCs, the overall features of the HC database are conserved for the previously reported HC species, in particular highlighting a set of HC species with clear high affinities for the α-(H) or β-(E) states, that will be more particularly analyzed in this study. Moreover, by increasing the number of HC species for which statistics are available, HCDB v2 made it possible to increase more than two folds H/E HC species with strong affinities for the α-helix (H) or β-strand (E) states (192 instead of 78). HCDB v2 contains information for all the existing species up to length 7, except one. Indeed, at this length, only one species (1111111, Peitch code P-127) is represented with less than 30 members). In a general way and consistent with previous observations [6], long hydrophobic runs are significantly under-represented in globular domains. This may be explained by the need of avoiding aggregation of partially folded intermediates. From length 8 onwards, some theoretical HC species are not present, with HC species of lengths 10, 11 and 12 with more than 30 members representing approximately 50%, 25% and 10% of the total number of theoretical HC species (**Figure 4C**). HCDB v2 thus includes the most frequent HC species associated with RSS within protein globular domains, gathering more than 4/5 of the total number of HCs within globular domains.

3.2. DISCORDANCE VERSUS CONCORDANCE

Considering the 180 HC species with affinities for α-helices (H/h) and β-strands (E/e) stored in HCDB v1, we have recently demonstrated at a large scale the leading role of the sequence periodicity of polar and non-polar amino acids (binary patterns) in the formation of RSS [15]. In particular, by breaking down binary patterns into their four basic units (called the Quarks (see before and **Figure 1**), we evidenced that the 88 HC species with strong affinities for RSS (H and E species) have clear binary patterns, with periodicities typical of α-helices (with a majority of *u* and *d* Quarks) and β-strands (with

a majority of *v* and *m* Quarks), respectively (see examples of two such clusters in **Figure 1**). In this context, HCs belonging to HC species with a high SS affinity that individually harbour a discordant behaviour, *i.e.* have an observed SS different from that expected from their binary pattern, must necessarily have one or more hydrophobic amino acids exposed to the solvent, which may participate in an interaction site (**Figure 1**).

To test this hypothesis, we thus evaluated to what extent discordant (abbreviated "**D**" for *Disagreeing with the expected behavior*) HCs with "strong" binary patterns (namely the 192 HC species with high RSS affinities (E and H) described in HCDB v2) have more exposed and interacting hydrophobic amino acids than concordant (abbreviated "**A**" for *Agreeing with the expected behavior*) clusters. The concordance (A)/discordance (D) status was assigned as previously reported [15]. At an individual level, an HC is defined as concordant (A) ("80%" rule) if at least 80% of its positions are observed in the HC species major state (as now reported in HCDB v2), as discordant (D) if less than 20% are observed in this same HC species major state, and as intermediate (I) otherwise (**Figures 2D and 3B**). One should expect, at global level (thus at the HC species level), that the number of concordant (A) HCs is higher than the number of discordant (D) HCs. This is what is actually observed, with 170 H or E HC species (out of 192) for which more concordant (A) HCs are observed than discordant (D) ones (**Table 1** and **Figure 3A** for the particular example of HC species of P-code 85). These HC species are described as "strong" (S) (**Figure 2E**). However, a few HC species, described as "weak" (W), do not obey the rule, with more discordant (D) HCs than concordant (A) ones (**Figure 2E**). This is in particular the case for HC species with low h and e affinities (which are however not considered further here) and for which almost equal (or at least not sufficiently different) numbers of members are found in the two states. More amazing are the rare cases of weak HC species with strong RSS affinities. Of note is that 21 of the 22 weak HC species have E affinities, whereas only one (P-2229, 100010110101) has an H affinity. These peculiarities may originate in part from the differences between the HC SS assignment and concordance (A)/discordance (D) assignment rules (OPS versus 80 % rules, respectively). This is for instance the case of HC species of P-codes P-169 (10101001) and P-113 (111001), having E affinities because found associated more frequently with β-strands than with α-helices, according to the OPS rule (**Figure 3A**). However, the HCs in those HC species are frequently

observed in an intermediate (I) state (**Figure 3B**), *i.e.* having between 20% and 80% of their positions assigned to E (for the two chosen examples, HC species P-113 has 348 intermediate (I) clusters versus 165 concordant (A) ones, HC species P-169 has 340 intermediate (I) clusters versus 74 concordant (A) ones). Accordingly, lower RSS coverage values (cov) are observed (66% and 67% for the chosen examples) and these clusters are more rarely found with at least 80% of their positions in the β-state, being therefore globally discordant (D). This means that in these particular cases, the HC limits are on average greater than those of the individual RSS. Of note is that the concerned HC species have generally mixed composition in Quark codes, being made of Quarks codes (or Q-codes) typical of both α-helices (*u* = 1001 and *d* = 10001) and β-strands (*v* = 11 and *m* = 101). For instance, species P-169 corresponds to the Q-code combination *mmu* and species P-113 to *vvu*. This general trend is observed for all the "weak" HC species with β-strand affinities (see the Quark composition of E/e weak (W) HC species and the E/e "strong" (S) ones in **Figure 5**). For HC species with low helix affinity (hS and hW), no clear difference could be highlighted. Note that the comparison of HC species with high helix affinity is not relevant, as there is only one HW HC species (see above P-2229, not represented in **Figure 5**)). The Q-code combination of this HC species (*dmvmm*) is however representative of a mixed behaviour.

3.3. SOLVENT EXPOSURE AND INTERACTION OF HYDROPHOBIC AMINO ACIDS

The solvent accessibilities of hydrophobic amino acids belonging to HC species with high RSS affinities (including or not the 22 weak HC species) were calculated in two different ways. The first way considers the relative solvent accessibility of the amino acids (a 36 % threshold was chosen according to Rost and Sander [22]), the second one a fixed, minimal absolute value of 40 $Å^2$, not depending on the considered amino acid (the 40 $Å^2$ value corresponds to 36 % of relative accessibility for alanine). We also considered the information included in the PDBSUM database about the interactions that those hydrophobic amino acids make with partners (interactions were distinguished between proteins, metals, ligands and nucleic acids). Solvent accessibility calculations were made by excluding all the 3D structure partners (protein, ligand, nucleic acid). **Table 2A**, considering strong (S/s) and weak (W/w) HC species with high affinity for α -helices (H) and β-strands (E), shows that there are more than twice as many HCs which have one or several hydrophobic amino acids exposed to solvent in a discordant (D) state than in a

concordant (A) one, when the relative solvent accessibility is considered (1.7 fold when a minimal absolute value of 40 $Å^2$ is considered). The same trend is observed when considering interactions that hydrophobic amino acids make with their partners, even though in the case of relative accessibility, the percentages are lower than in the case of absolute accessibility. We also considered only lateral chains for the calculation of interacting amino acids (instead of the whole set of atoms), but this did not give significantly different results (data not shown). No significant difference was also observed when considering separately HC species with H and E affinities (**Table 2B**). The same calculations were finally made by considering only strong (S/s) HC species (**Table 2C**), but this led to only slight differences.

3.4. Analyzing 3D structures of interacting clusters with concordant/discordant behaviors

We first examined examples of 3D structures of **concordant (A) HCs with interacting hydrophobic amino acids** (as reported in PDBSUM), as these are more present than first expected from the binary pattern preference rule (23.7 % of hydrophobic clusters, using the 36 % rule, **Table 2A**). Two first examples are concordant HCs for which interacting hydrophobic amino acids are exposed within a hydrophobic crevice (frequenin, **Figure 6A**) or a cradle (GAF domain of a cyanobacteriochrome, **Figure 6B**). In both cases, ligands shield the crevice/cradle from the solvent. The crevice of frequenin, a member of the myristoyl-switch Ca^{2+}-binding proteins, is indeed involved in ligand/membrane recognition (here mimicked by a polyethelyne glycol fragment used for crystallization [23]), whereas the cradle of the cyanobacteriochrome covalently links phycoviolobilin [24]. These internal cavities, almost permanently shielded from solvent, may thus be considered as "masked" protein hydrophobic cores. A similar situation is observed with the cavities of heme-binding proteins, among which haemoglobins (highlighted within the 40 $Å^2$ database, data not shown). Other frequently encountered examples are multimeric architectures (mostly dimers) which form permanent or obligate interactions. The first example of such a case is reported in **Figure 6C**, with the (alpha-beta) heterodimer formed by a phycobiliprotein, a component of the phycobilisomes. Here, the exposed/interacting hydrophobic amino acids within an α-helix participate in the hydrophobic core formed inter-molecularly. The second example (**Figure 6D**) is found within the multimeric architecture formed by the T-cell receptor

IG-like domains, in which the exposed/interacting hydrophobic amino acids within a β-strand also participate in the hydrophobic core formed inter-molecularly. More complex cases of obligate interactions are those of intertwined subunits, as found for instance in repressor proteins (**Figure 6E**). Finally, there are also cases, more difficult to interpret, of exposed hydrophobic amino acids at the end of regular secondary structures (thus with an HC larger than the RSS), already included within loops participating in the protein interfaces (**Figure 6F**). In conclusion, in most of the examined examples of concordant (A) clusters, hydrophobic amino acids exposed and interacting with partners participate in protein hydrophobic cores, formed either inter-molecularly (obligate interactions, including secondary structure swapping) or intra-molecularly (ligand-binding crevices or internal binding sites).

We next examined the 3D structures of **discordant (D) HCs with exposed hydrophobic amino acids**, for which interaction(s) are also reported in PDBSUM. There are many cases of coil (C) positions observed where regular secondary structures (α-helices or and β-strands) are expected. These correspond to true coils (some of which participating in active sites, **Figure 7A**) but also sometimes positions which are misassigned by secondary structure assignment methods (irregular helices or extended segments, data not shown). There are also many cases of β-strands observed instead of expected α-helices and α-helices observed instead of expected β-strands (**Figures 7B to 7F**). These are found in interfaces formed by heterodimers (**Figures 7B** and **7E**) or homodimers (**Figure 7D**). Cases of intertwined homodimers are also encountered, corresponding to obligate interactions (**Figures 7C** and **7F**). In conclusion, in most of the examined examples of discordant (D) HCs, hydrophobic amino acids exposed and interacting with partners participate in protein/protein or protein/ligand interfaces.

3.5. AMINO ACID SEQUENCE COMPOSITION OF CONCORDANT/DISCORDANT CLUSTERS

Previous observations based on HCDB v1 indicated that amino acid composition is critical for distinguishing the concordant (A)/discordant (D) behavior of a given HC relatively to its HC species affinity [15]. Amino acids sequence profiles are given here for each HC species of HCDB v2 (**http://osbornite.impmc.upmc.fr/HCADB/dictionary**), displaying clear differences between the concordant (A)/discordant (D) states. Hence, for the example shown in **Figure 3A** (P-code 85), the amino acids composition in the prevalent

secondary structure state (E) is clearly different from those observed in other states (H, M or C) for both hydrophobic and non-hydrophobic positions. The amino acids profiles of the concordant (A) and discordant (D) states are logically in agreement with those of major (E) and minor (H) RSS states respectively observed for this HC species. Of note is that non-hydrophobic positions of discordant (D) clusters are often occupied by non-polar amino acids (alanine in the case of H assignments) or polar amino acids that mask their polarity through H-bonds (threonine and serine in the case of E assignments). These amino acids can thus well adapt the protein cores, as depicted in **Figure 1**. On a global level, for all the species, we also computed propensities of the different amino acids for the concordant (A) and discordant (D) states for each affinity (considering only H and E states) (**Figure 8**). We grouped amino acids according to their general propensities towards secondary structures, as reported in [3], thereby distinguishing α-helix (H), β-strand (E) and coil (C) forming residues. Of note is the neutral behavior of histidine towards any secondary structure. As noticed previously, based on HCDB v1 [15], we observe that: (i) **for HC species typical of α-helices (H - red)**, the propensities for α-helix forming amino acids (alanine, glutamic acid, glutamine, lysine, arginine, leucine, methionine) decreased in the discordant state for the benefit of loop-forming (glycine, serine) or β-strand-forming (valine, threonine, cysteine, tyrosine, phenylalanine, tryptophane) residues, (ii) **for HC species typical of β-strand (E - green)**, the propensities for β-strand-forming residues (valine, isoleucine, threonine, cysteine) diminished in the discordant state for the benefit of loop-forming residues (glycine, aspartic acid, asparagine, serine) and α-helix-forming residues (alanine, glutamic acid, glutamine, leucine, methionine). Of note was that H propensities of methionine and leucine for the H and E concordant (A) states are relatively neutral and similar, whereas the E propensities of valine, isoleucine, phenylalanine, tyrosine and tryptophane associated with the concordant (A) state are well pronounced and clearly different from the H propensities. This highlighted the fact that β-strand-forming hydrophobic amino acids play prominent role in the formation of β-strands, whereas this role is played in particular by non-hydrophobic amino acids in α-helices, especially through intra-helical bonds. The α-helix forming amino acids (alanine, glutamic acid, glutamine, lysine, arginine, leucine, methionine) are more associated with concordant H HC species and the β-strand-forming residues (valine, isoleucine, threonine, cysteine) are more associated with concordant (A) E HC species. Interestingly, the highest

propensities are observed in the discordant (D) state for the three aromatic amino acids (phenylalanine, tyrosine, tryptophane), indicating that there are more particularly involved in protein interfaces, for HC species with affinities for both the α and β states. Such role is supported by propensities of amino acids to be exposed or interacting within concordant (A) and discordant (D) clusters with E affinity and within discordant (D) clusters with H affinity (**Supplementary Data 3**). Finally, higher propensities in the discordant (D) state with both α-helix and β-strand forming HCs are also logically observed for the SS indifferent histidine (H), as well as the loop-forming residues (glycine, serine; proline is not considered as not included in HC), except from aspartic acid which behaves as an α-helix forming amino acid. On another hand, according to their particular behaviour and low RSS coverage values, we observed that propensities for coil-forming amino acids and for cysteine and threonine are higher in the weak HC species (EW), opposite to HS and ES strong cluster species, for which propensities for α-helix- and β-strand-forming amino acids are respectively higher (**Supplementary Data 4**). In conclusion, considering the amino acid composition of an HC relative to the profiles described for each HC species should help to distinguish between concordant (A) and discordant (D) behaviors.

4. DISCUSSION

In this study, we first provide an updated Hydrophobic Cluster DataBase (HCDBv2), almost doubling the number of populated species relative to HCDBv1 [14]. HCDBv2, covering more than 4/5 of the total number of HCs encountered in protein globular domains, gives information especially on species with high affinities for α-helices (H) or β-strands (E), which we often observe to constitute the signatures of folds, participating in their hydrophobic core and being much more conserved than sequences within families of homologous sequences [25] (**Figure 9**). Therefore, considering these HCs can be useful to highlight remote relationships at high level of sequence divergence ([3] and [26-31] for some examples), as they vary around preferential HC substitution schemes. Such HCs, participating in the hydrophobic core, generally adopt secondary structures corresponding to the affinity of the HC species to which they belong (green letters in **Figure 9**). Here, we have focused our analysis on the particular case of HCs disobeying this binary pattern preference (and called therefore discordant HCs) and showed, based on a set of representative 3D structures extracted from the SCOP database, that such

sequences have hydrophobic amino acids more often exposed to solvent and more interacting with partners. This relatively simple approach may thus be of potential interest for highlighting hydrophobic components of protein interfaces, based on the only knowledge of a single amino acid sequence. Among the different factors at play, hydrophobicity has already been recognized as a key factor and surface hydrophobicity can be used to identify regions of a protein's surface most likely interacting with a binding partner [32, 33]. Hence, clusters of exposed amino acids with the highest hydrophobicity scores are usually found to include the attachment site for the complexed molecules [32]. In the context of the huge amount of sequences for which experimental 3D structures are not solved, deriving some information for distinguishing, from the only knowledge of a single amino acid sequence, hydrophobic amino acids which are likely exposed to solvent (and possibly interacting with partners) from those participating in the hydrophobic core of globular domains (participating in either intra- or inter-molecular (obligate) interactions), may thus be of potential interest.

Amino acid profiles derived from each HC species in the different states, reported here in details, can be used to highlight those HCs that probably deviate from concordant behavior. Further analyses are being carried out to develop an automatic prediction tool for prediction the concordant/discordant state of clusters using for instance machine learning approaches (data not shown), which however must overcome the limited sizes of the considered sequence segments, and also take into account the flanking sequences and non local interactions.

It can be noted that some positions within the binary patterns can sometimes be occupied predominantly by mimetic amino acids, *i.e.* amino acids (such as alanine, threonine, serine or cysteine) which can substitute for the seven strong hydrophobic amino acids that are currently used in the HCA alphabet [12]. This is especially the case for discordant HCs, for which this mimetic behavior, if integrated within the hydrophobic alphabet, makes the HC shifting towards species with opposite affinity (for instance an alanine at position 3 (often observed in discordant clusters) of species P-19 (10011), with H affinity leads to shift towards species P-23 (10111) with E affinity). The approach could thus be further improved by considering a three-state binary pattern, splitting the "0" class in two sub-classes, one of which corresponding to amino acids with possible hydrophobic behavior.

Worth noting is that discrepancies between predicted (expected) and experimentally determined (observed) secondary structures may also be associated with fold-switching proteins, characterized by regions able to adopt two very different secondary structures [34, 35]. Cases of structural ambivalence are also observed in some amyloid-forming proteins undergoing α-to-β conformational conversions (**Supplementary Data 5**). In such cases, the predicted secondary structure state generally corresponds to that observed in one of the conformational states, associated with the lowest energy. A discordant behavior of a HC relative to its HC species preference may thus be considered with a possibility of switching towards a concordant state.

Finally, it should be stressed that the information gained here on the fundamental bricks of protein globular domains can also be considered to highlight regions within disordered segments (called intrinsically disordered regions (IDRs)) which are likely to fold in contact with partners. We have previously shown that HCA, and the derived SEG-HCA tool, are particularly well adapted to identify foldable domains within proteins [36, 37]. In the case of IDRs, these foldable domains are often limited to one or two HCs, with strong RSS affinity [38]. In light of known 3D structure of complexes in which IDRs participate (exemplified here with a SUMO-interacting motif (SIF) present in human MCAF1 – **Supplementary Data 6**), the observed RSS is often in agreement with the HC species affinity. This means that hydrophobic amino acids of the HC interact with hydrophobic patches present at the partner surface, thereby participating in the hydrophobic core of the partner. Using HCDBv2 can thus also give useful information to predict the behavior of such foldable segments from IDPs, in addition to that of well-folded globular domains.

Acknowledgments

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

Authors contributions

A.L. built HCDB v2. A.L., T.B.-F., J.R. and F.Q. carried out the theoretical work (conceptualization, data curation and analysis) relative to concordant/discordant clusters. I.C., F.S. and J.P.M. were involved in the conceptualization, design of the study and analysis of the results. I.C. initiated and supervised the project. A.L., T.B.-F., J.R and I.C wrote the paper. All authors discussed the results and commented on the manuscript.

References

[1] K. Dill, Dominant forces in protein folding, Biochemistry, 29 (1990) 7133-7155.

[2] T. Creighton, Proteins: Structures and molecular properties., 2nd Ed ed., Freeman Press, New York, 1993.

[3] I. Callebaut, G. Labesse, P. Durand, A. Poupon, L. Canard, J. Chomilier, B. Henrissat, J.P. Mornon, Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives, Cell Mol Life Sci, 53 (1997) 621-645.

[4] P. Huang, S. Boyken, A. Baker, The coming of age of *de novo* protein design, Nature, 537 (2016) 320-327.

[5] G. Bellesia, A.I. Jewett, J.E. Shea, Sequence periodicity and secondary structure propensity in model proteins, Protein Sci, 19 (2010) 141-154.

[6] R. Schwartz, S. Istrail, J. King, Frequencies of amino acid strings in globular protein sequences indicate suppression of blocks of consecutive hydrophobic residues, Protein Sci, 10 (2001) 1023-1031.

[7] M.W. West, M.H. Hecht, Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins, Protein Sci, 4 (1995) 2032-2039.

[8] H. Xiong, B.L. Buckwalter, H.M. Shieh, M.H. Hecht, Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides, Proc Natl Acad Sci U S A, 92 (1995) 6349-6353.

[9] S. Ventura, L. Serrano, Designing proteins from the inside out, Proteins, 56 (2004) 1-10.

[10] N. Koga, R. Tatsumi-Koga, G. Liu, R. Xiao, T. Acton, G. Montelione, B. D, Principles for designing ideal protein structures, Nature, 491 (2012) 222-227.

[11] C. Gaboriaud, V. Bissery, T. Benchetrit, J.P. Mornon, Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences, FEBS Lett., 224 (1987) 149-155.

[12] S. Woodcock, J.P. Mornon, B. Henrissat, Detection of secondary structure elements in proteins by hydrophobic cluster analysis, Protein Eng, 5 (1992) 629-635.

[13] J. Hennetin, K. Le Tuan, L. Canard, Colloc'h, N, J.P. Mornon, I. Callebaut, Non-intertwined binary patterns of hydrophobic/nonhydrophobic amino acids are considerably better markers of regular secondary structures than nonconstrained patterns, Proteins, 51 (2003) 236-244.

[14] R. Eudes, K. Le Tuan, J. Delettre, J.P. Mornon, I. Callebaut, A generalized analysis of hydrophobic and loop clusters within globular protein sequences, BMC structural biology, 7 (2007) 2.

[15] J. Rebehmed, F. Quintus, J.P. Mornon, I. Callebaut, The respective roles of polar/nonpolar binary patterns and amino acid composition in protein regular secondary structures explored exhaustively using hydrophobic cluster analysis, Proteins, 84 (2016) 624-638.

[16] N. Fox, S. Brenner, J. Candonia, SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures, Nucleic Acids Res, 42 (2014) D304-309.

[17] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers, 22 (1983) 2577-2637.

[18] D. Frishman, P. Argos, Knowledge-based protein secondary structure assignment., Proteins, 23 (1995) 566-579.

[19] A.V. Finkelstein, N.S. Bogatyreva, S.O. Garbuzynskiy, Restrictions to protein folding determined by the protein size, FEBS Lett, 587 (2013) 1884-1890.
[20] M. Levitt, Nature of the protein universe, Proc. Natl. Acad. Sci. USA, 106 (2009) 11079–11084.
[21] A. Scaiewicz, M. Levitt, The language of the protein universe, Curr Opin Genet Dev, 35 (2015) 50-56.
[22] B. Rost, C. Sander, Conservation and prediction of solvent accessibility in protein families, Proteins, 20 (1994) 216-226.
[23] Y. Bourne, J. Dannenberg, V. Pollmann, P. Marchot, O. Pongs, Immunocytochemical localization and crystal structure of human frequenin (neuronal calcium sensor 1), J Biol Chem, 276 (2001) 11949-11955.
[24] E. Burgie, J. Walker, G. Phillips, R. Vierstra, A photo-labile thioether linkage to phycoviolobilin provides the foundation for the blue/green photocycles in DXCF-cyanobacteriochromes, Structure, 21 (2013) 88-97.
[25] A. Poupon, J.P. Mornon, Populations of hydrophobic amino acids within protein globular domains: identification of conserved "topohydrophobic" positions, Proteins, 33 (1998) 329-342.
[26] I. Callebaut, J.C. Courvalin, J.P. Mornon, The BAH (bromo-adjacent homology) domain: a link between DNA methylation, replication and transcriptional regulation., FEBS Lett, 446 (1999) 189-193.
[27] I. Callebaut, J. de Gunzburg, B. Goud, J. Mornon, RUN domains: a new family of domains involved in Ras-like GTPase signaling, Trends Biochem Sci, 26 (2001) 79-83.
[28] I. Callebaut, J.P. Mornon, From BRCA1 to RAP1: a widespread BRCT module closely associated with DNA repair, FEBS Lett, 400 (1997) 25-30.
[29] I. Callebaut, J.P. Mornon, OCRE: a novel domain made of imperfect, aromatic-rich octamer repeats, Bioinformatics, 21 (2005) 699-702.
[30] I. Callebaut, J.P Mornon, LOTUS, a new domain associated with small RNA pathways in the germline., Bioinformatics, 26 (2010) 1140-1144.
[31] I. Callebaut, D. Moshous, J.P. Mornon, J.P. de Villartay, Metallo-beta-lactamase fold within nucleic acids processing enzymes: the beta-CASP family, Nucleic Acids Res, 30 (2002) 3592-3601.
[32] C.-J. Tsai, R. Nussinov, Hydrophobic folding units at protein-protein interfaces. Implications to protein folding and to protein-protein association, Protein Sci, 6 (1997) 1426-1427.
[33] L. Young, R. Jernigan, D. Covell, A role for surface hydrophobicity in protein-protein recognition, Protein Sci, 3 (1994) 717-729.
[34] M. Young, K. Kirshenbaum, K.A. Dill, S. Highsmith, Predicting conformational switches in proteins, Protein Sci, 8 (1999) 1752-1764.
[35] L. Porter, L. Looger, Extant fold-switching proteins are widespread, Proc Natl Acad Sci U S A, 115 (2018) 5968-5973.
[36] G. Faure, I. Callebaut, A comprehensive repertoire of foldable segments within genomes, PLoS Comput Biol, 9 (2013) e1003280.
[37] T. Bitard-Feildel, I. Callebaut, Exploring the dark foldable proteome by considering hydrophobic amino acids topology, Sci Rep, 7 (2017) 41425.
[38] T. Bitard-Feildel, A. Lamiable, J.P. Mornon, I. Callebaut, Order in disorder as observed by the "hydrophobic cluster analysis" of protein sequences. Proteomics, in press (2018).

[39] K. Brejc, R. Ficner, R. Huber, S. Steinbacher, Isolation, crystallization, crystal structure analysis and refinement of allophycocyanin from the cyanobacterium Spirulina platensis at 2.3 Å resolution , J Mol Biol 249 (1995) 424-440.
[40] G. Stewart-Jones, A. McMichael, J. Bell, D. Stuart, E. Jones, A structural basis for immunodominant human T cell receptor recognition, Nat Immuno, 4 (2003) 657-663.
[41] J. Schildbach, A. Karzai, B. Raumann, R. Sauer, Origins of DNA-binding specificity: role of protein contacts with the DNA backbone, Proc Natl Acad Sci U S A, 96 (1999) 811-817.
[42] V. Vandavasi, K. Taylor-Creel, R. McFeeters, L. Coates, H. McFeeters, Recombinant production, crystallization and X-ray crystallographic structure determination of peptidyl-tRNA hydrolase from Salmonella typhimurium, Acta Crystallogr F Struct Biol Commun, 70 (2014) 872-877.
[43] B.S. Rajagopal, A.N. Edzuma, M.A. Hough, K.L. Blundell, V.E. Kagan, A.A. Kapralov, L.A. Fraser, J.N. Butt, G.G. Silkstone, M.T. Wilson, D.A. Svistunenko, J.A.R. Worrall, The hydrogen-peroxide-induced radical behaviour in human cytochrome c-phospholipid complexes: implications for the enhanced pro-apoptotic activity of the G41S mutant, Biochem J, 456 (2013) 441-452.
[44] L. Mazzarella, G. Bonomi, M. Lubrano, A. Merlino, A. Riccio, A. Vergara, L. Vitagliano, C. Verde, G. di Prisco, Minimal structural requirements for root effect: crystal structure of the cathodic hemoglobin isolated from the antarctic fish *Trematomus newnesi*, Proteins, 62 (2006) 316-321.
[45] A.M. Haapalainen, M.K. Koski, Y.M. Qin, J.K. Hiltunen, T. Glumoff, Binary structure of the two-domain (3R)-hydroxyacyl-CoA dehydrogenase from rat peroxisomal multifunctional enzyme type 2 at 2.38 Å resolution, Structure, 11 (2003) 87.
[46] P. Madoori, H. Agustiandari, A. Driessen, A. Thunnissen, Structure of the transcriptional regulator LmrR and its mechanism of multidrug recognition, EMBO J, 28 (2009) 156-166.
[47] V. Bamford, S. Bruno, T. Rasmussen, C. Appia-Ayme, M. Cheesman, B. Berks, A. Hemmings, Structural basis for the oxidation of thiosulfate by a sulfur cycle enzyme., EMBO J, 21 (2002) 5599-55610.
[48] J. Ren, S. Sainsbury, S. Combs, R. Capper, P. Jordan, N. Berrow, D. Stammers, N. Saunders, R. Owens, The structure and transcriptional analysis of a global regulator from *Neisseria meningitidis*, J Biol Chem, 282 (2007) 14655-14664.

Tables

Status	Number of cluster species	Percentage	Number of clusters
Strong S	139	72.40	51946
strong s	31	16.15	13939
Weak W	18	9.38	3090
weak w	4	2.08	252

Table 1 : Strong (S/s)/weak (W/w)status of the 192 HC species with high RSS affinity (H and E). If a HC species contains N_A concordant (A) HCs, N_D discordant (D) ones and N_I intermediate (I) ones, then it is considered "strong" if $N_A > 0.8\ N_D$ (N_I is not taken into account), highlighting that the behavior of individual HCs is generally in agreement with the HC species affinity. Otherwise, it is considered as "weak" (more individual cases of discordant (D) than concordant (A) HCs). Uppercase/lowercase letters indicate high and low tendencies, *i.e.* differences between the N_A and N_D count greater/lower than 20% of the total of concordant (A) /discordant (D) HCs (see Material and Methods).

A)

	Exposed (> 36 %)			Exposed (> 40 Å²)		
HC state	**A - concordant**	**D - discordant**	**I – intermediate**	**A - concordant**	**D - discordant**	**I – intermediate**
Total	35288	16758	15779	35288	16758	15779
Exposed	8369	9337	6709	15369	12480	9841
Exposed %	**23.7**	**55.7**	**42.6**	**43.6**	**74.5**	**62.4**
Interacting	2930	3042	1975	4601	4021	2768
Interacting %	**8.3**	**18.2**	**12.5**	**13.0**	**24.0**	**17.5**
Protein	*2548*	*2546*	*1533*	*3693*	*3114*	*1942*
Metal	*11*	*31*	*33*	*32*	*51*	*56*
Ligand	*386*	*476*	*419*	*919*	*895*	*820*
Nucleic acid	*10*	*27*	*15*	*37*	*44*	*26*

B)

Accessibility	**Exposed (> 36 %) H**			**Exposed (> 36 %) E**		
HC state	**A - concordant**	**D - discordant**	**I – intermediate**	**A - concordant**	**D - discordant**	**I – intermediate**
Total	14319	6462	5182	22862	13103	14318
Exposed	3832	3354	2434	4863	7462	5856
Exposed %	**26.8**	**51.9**	**47.0**	**21.3**	**56.9**	**40.9**
Interacting	1255	1073	788	1809	2497	1633
Interacting %	**8.8**	**16.6**	**15.2**	**7.9**	**19.1**	**11.4**

C)

Accessibility	**Exposed (> 36 %) H**			**Exposed (> 36 %) E**		
HC state	**A - concordant**	**D - discordant**	**I – intermediate**	**A - concordant**	**D - discordant**	**I – intermediate**
Total	12064	4617	4411	16803	6809	6882
Exposed	3330	2446	2145	3285	4005	2761
Exposed %	**27.6**	**53.0**	**48.6**	**19.6**	**58.8**	**40.1**
Interacting	1080	814	707	1237	1287	704
Interacting %	**9.0**	**17.6**	**16.0**	**7.4**	**18.9**	**10.2**

Table 2: Solvent exposure and interaction of hydrophobic amino acids

A) HCs included in HC species with H and E affinities, in concordant (A), discordant (D) or intermediate (I) states: total number of HCs, number of HCs with a least one hydrophobic amino acid exposed to solvent and number of HCs with at least one hydrophobic amino acid in interaction with partners (protein, metal, ligand or nucleic acids), as reported in PDBSUM.

B) Number of HCs in the different states, separated following the species affinities (H or E).

C) Same as the above, but taking into account only the strong HC species.

Figures

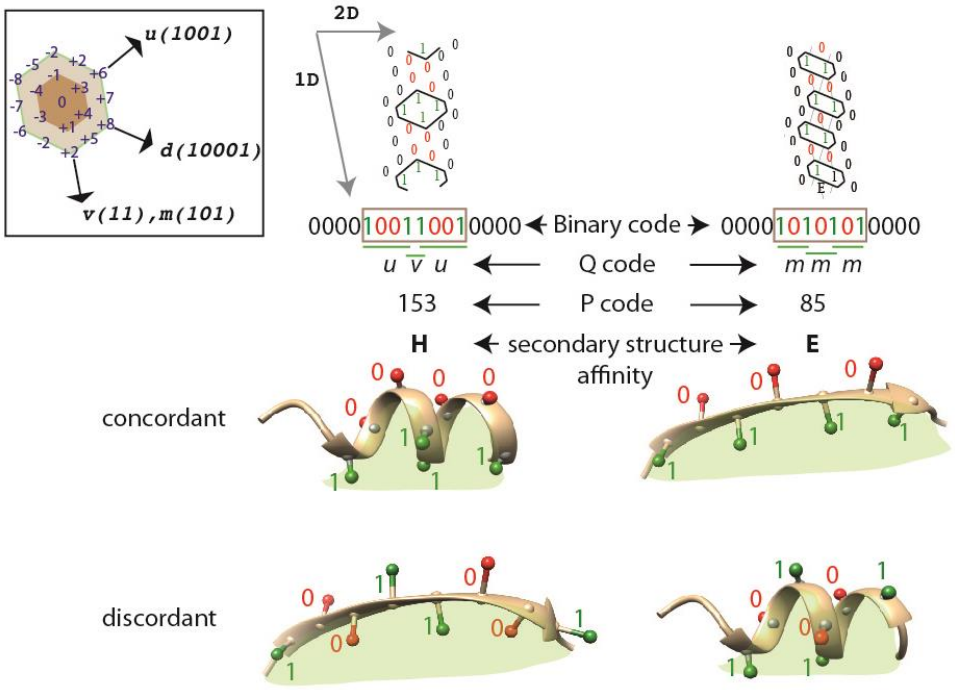


Figure 1: Binary patterns and secondary structure preferences. The binary patterns (*patterns of hydrophobic (coded 1) and non-hydrophobic (coded 0) amino acids*) considered here are constrained binary patterns (or hydrophobic clusters), which have been shown to well match the limits of regular secondary structures (RSSs) [12]. They differ from simple binary patterns through the consideration of a connectivity distance, linked to a bidimensional (2D) representation of the sequence (shown on top). On this 2D representation, the sequence is written on a duplicated α-helical net, and the strong hydrophobic amino acids (V, I, L, M, F, Y, W) are encircled, forming hydrophobic clusters (HCs). The connectivity distance corresponds to the minimal number (4 in the case of an α-helical net) of non-hydrophobic amino acids necessary to split a binary pattern into two separate hydrophobic clusters (HCs) (black 0, in contrast to red 0 belonging to the HCs). HC species are designated with their binary, Quark (Q)- and Peitsch (P)-codes. In the binary code, 1 stands for any strong hydrophobic amino acid (V I L F M Y W) and 0 for any other amino acids except from proline (P)). In the Q-code (also defined on the 2D net shown at left), *v* (11) stands for *"vertical*," *m* (101) for "mosaic", *u* (1001) for "*up*" and *d* (10001) for "*down*". The P-code is defined as the sum of the powers of 2, indexed according to the position of each number of the binary code (the last position corresponds to 0), each power being multiplied by the binary code value (*e.g* for the HC species of binary code **1**0**1**0**1**0**1**, the P-code is **1**x2^6+0x2^5+**1**x2^4+0x2^3+**1**x2^2+0x2^1+**1**x2^0= 64+0+16+0+4+0+1=85). The two shown examples are cluster species highly associated with α-helices (P-153) and β-strands (P-85), according to HCBD v2 (**Figure 3A**).
In concordant (A) HCs (*i.e.* HCs for which the observed SS state is in agreement with the behavior expected from their binary pattern), all the hydrophobic positions (coded 1 and highlighted in green on the 3D representation (only Cβ shown)) are buried within the hydrophobic core (shaded green). In discordant (D) HCs, some of the hydrophobic positions are exposed to solvent.

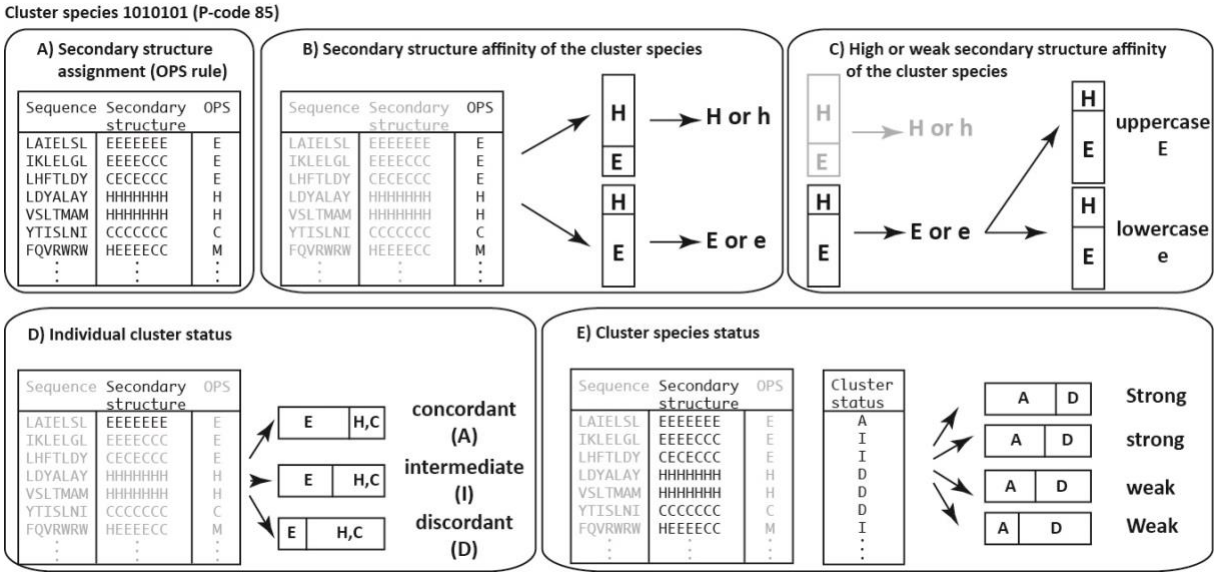
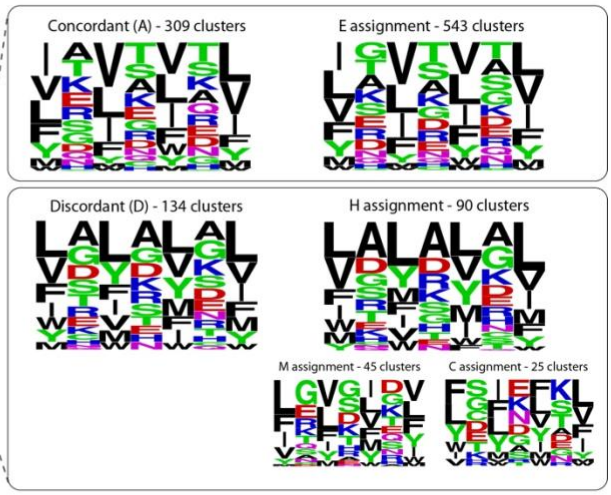


Figure 2: General workflow followed in this study. A) Secondary structures (SS) were assigned per amino acid position of each 3D structure using DSSP or STRIDE (only results with DSSP assignments are presented throughout the manuscript). Then, SSs associated with the considered hydrophobic clusters (HCs) are assigned using the OPS rule ("One Position is Sufficient"), assuming that HC limits are generally well covered by the regular secondary structures (RSS) (as assessed by the mean coverage (cov) calculated by HC species, see **Figure 3A**). H stands for α-helix, E stands for β-strand, C for coil and M for multiple (both H and E assignments found within the HC limits). **B)** The HC species SS affinity is determined by considering the highest propensity for one SS state**. C)** High (upper case) or low (lower case) SS affinity is defined, depending on the difference between the highest (first) and second SS propensities (see Material and Methods for the formula). **D)** At the individual level, the HC status indicates whether this one adopts or not the SS affinity observed for the HC species to which it belongs. Hence, a HC is defined as concordant (A *after "in Agreement with the expected behavior"*) or discordant (D *after "in Disagreement with the expected behavior"*) if at least 80 % or less than 20 % of the positions are assigned to the HC species affinity, intermediate (I) otherwise. **E)** The HC species status appreciates the level to which the individual HCs are globally concordant ("strong" species) or discordant ("weak" species). This status is refined in strong (upper case) and weak (lower case) behavior, depending on the difference between the numbers of concordant and discordant HCs within the HC species (see Material and Methods for the formula).

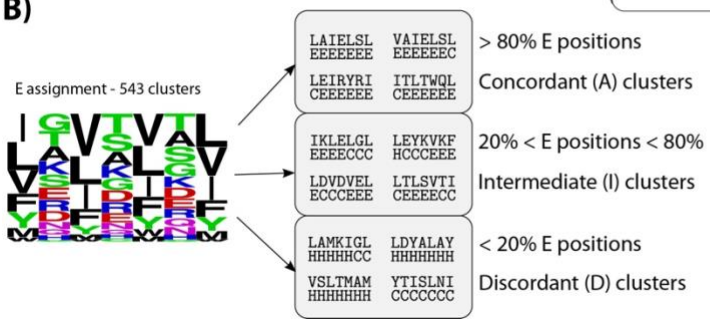
Figure 3: HCDB v2.

A)

					Secondary structures						
								DSSP			
								propensities			
P-code	Q-code	binary code	#	% id	affinity	status	cov	H	E	C	M
3	*v*	11	25554	95	h	W	90.10	0.57	0.54	0.48	0.06
5	*m*	101	14472	90	e	W	80.28	0.36	0.74	0.48	0.09
7	*vv*	111	10768	95	E	S	87.67	0.29	1.05	0.19	0.06
9	*u*	1001	8908	95	h	W	80.67	0.81	0.43	0.41	0.10
11	*mv*	1011	6544	95	e	s	80.18	0.45	0.96	0.16	0.09
13	*vm*	1101	6137	95	e	s	82.82	0.47	0.90	0.21	0.13
15	*vvv*	1111	3955	95	E	S	92.29	0.14	1.33	0.06	0.04
17	*d*	10001	5306	95	h	s	82.68	0.89	0.43	0.31	0.18
19	*uv*	10011	4460	95	H	S	89.84	1.06	0.41	0.13	0.14
21	*mm*	10101	3791	95	E	s	79.72	0.31	1.08	0.20	0.15
23	*mvv*	10111	2463	95	E	S	84.85	0.36	1.16	0.04	0.08
25	*vu*	11001	4489	95	H	s	84.99	1.01	0.45	0.14	0.16
27	*vmv*	11011	2524	95	h	s	90.73	0.87	0.64	0.06	0.14
29	*vvm*	11101	2203	95	E	S	84.65	0.35	1.14	0.07	0.12
31	*vvvv*	11111	806	90	E	S	92.01	0.21	1.32	0.02	0.09
...	...	...	...	...	...	...	...	...	...	...	...
85	***mmm***	**1010101**	**703**	**95**	**E**	**S**	**76.74**	**0.20**	**1.29**	**0.09**	**0.23**
113	*vvd*	1110001	794	95	E	W	65.70	0.38	1.03	0.12	0.35
153	*uvu*	10011001	1022	95	H	S	90.28	1.31	0.21	0.04	0.11
169	*mmu*	10101001	578	90	E	W	66.70	0.35	1.10	0.11	0.31



B)



A) Extract of the HCDB v2 (the full database can be found at **http://osbornite.impmc.upmc.fr/HCADB/dictionary**). Hydrophobic cluster (HC) species are designated with their Peitsch (P)- Quark (Q)- and binary codes, as detailed in the legend of **Figure 1**. For each HC species, the level of redundancy (% id) considered for the analysis is indicated, as well as the total number of HCs (occurrence: #), the propensities for the secondary structure classes, the resulting secondary structure affinity (E/e for strand, H/h for helix, M/m for multiple, with upper and lower cases representing high and low affinities, respectively) and global status (strong or weak according to the prevalence of concordant (A)/discordant (D) HCs relative to the affinity). The mean coverage (cov) of the HC species by the affinity SS is also indicated. SS assignment was made using DSSP and HC SS assignment was made following the OPS rule.

Are shown, for illustration purposes, the first HC species of HCDB v2, classified according to their length, as well as the two HC species with a global concordant behaviour (status S) and with high affinity for α-helices (H, P-153) and β-strands (E, P-85). Two examples of HC species with a high affinity for β-strands (E) but with a global discordant behaviour (status W) are also shown (P-113 and P-169). The Web-logos of HC P-85 are displayed at right, for the different SS and concordant (A) /discordant (D) assignments (see **Figure 2** for a detailed description of concordant (A) /discordant (D) states).

B) The concordant (A) /discordant (D) state of an individual HC cluster (illustrated here for the same HC species P-85, with six different amino acid sequences) is defined if at least 80 % or less than 20 % of the positions are assigned to the HC species affinity, respectively, intermediate (I) otherwise.

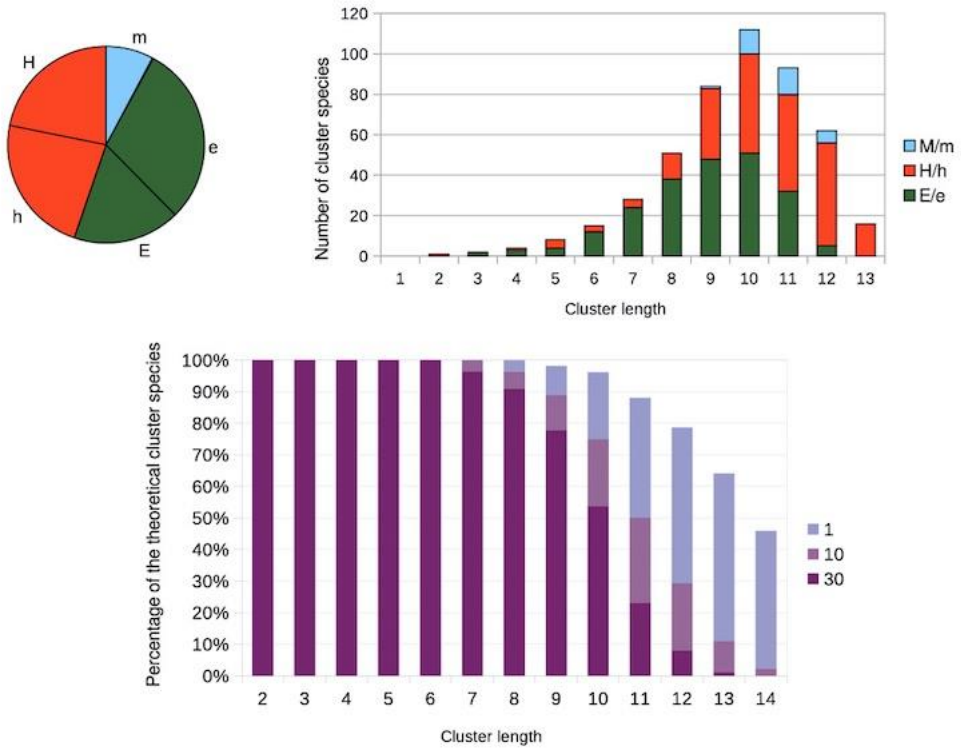


Figure 4: HCDB v2 main features

A) Global distribution of the HC species with at least 30 instances according to their SS affinities. H/h, E/e and M/m stand for helix, strand and multiple states, respectively, with upper and lower cases designating strong and low affinities. **B)** Distribution of the number of HC species, according to their lengths (in amino acids) and focused on cluster species with high affinities for the helix (H), strand (E) and multiple (M) states. **C)** Percentage of HC species with at least 30 cluster occurrences (dark purple), between 30 and 10 occurrences (purple) and only 1 occurrence (blue)) relative to the total number of theoretical hydrophobic cluster species present in the database, according to cluster length (in amino acids).

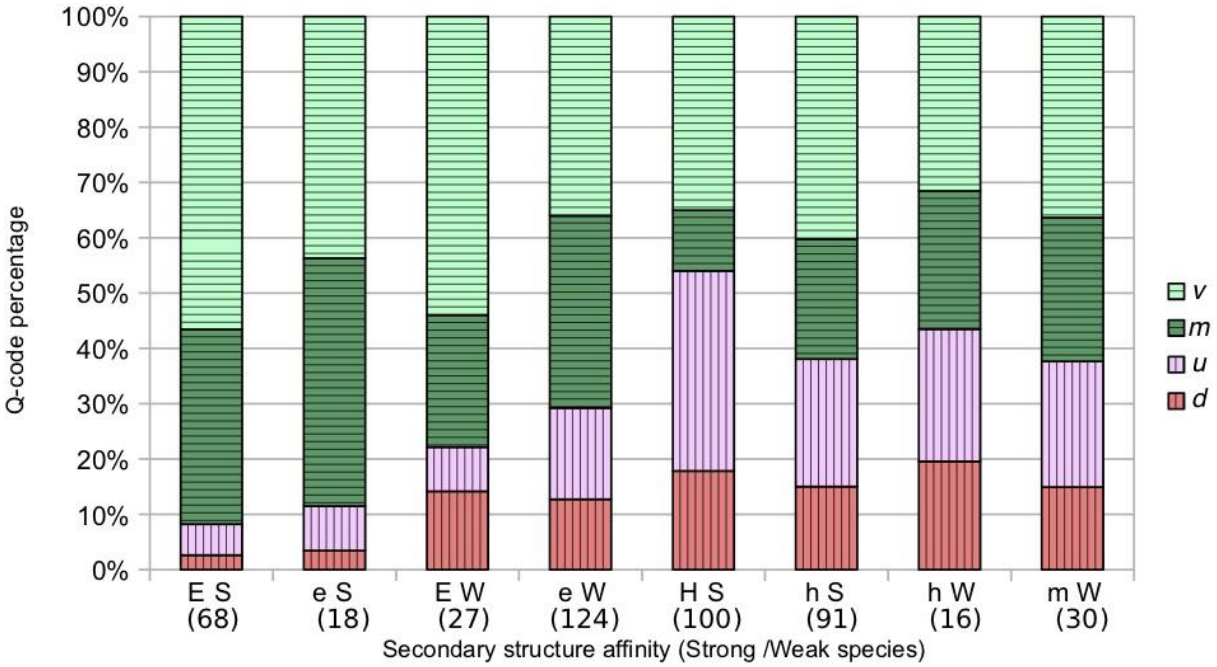


Figure 5: Composition in Quark (Q)-codes of HC species with strong and low helix (H/h) and strand (E/e) affinities, distinguishing global concordant (Strong (S))/discordant (Weak (W)) behaviors. *v* (11) stands for "vertical", *m* (101) for "mosaic", *u* (1001) for "up" and *d* (10001) for "down". *v* and *m* are Quark typical of β-strands (green), *u* and *d* of α-helices (red). *u* and *d* Q-codes are more frequent in E/e HC species with a discordant behavior (weak (W) HC species) than with a concordant one (strong (S) HC species), whereas there is no clear difference between strong (S) and weak (W) HC species with low helix affinity (h). Weak (W) HC species with high helix affinity (H W) are not represented, as there is only one case of such HC species (P-2229, Q-code DVMVV). The numbers in brackets under the X-axis indicate how many HC species of the given class were considered.

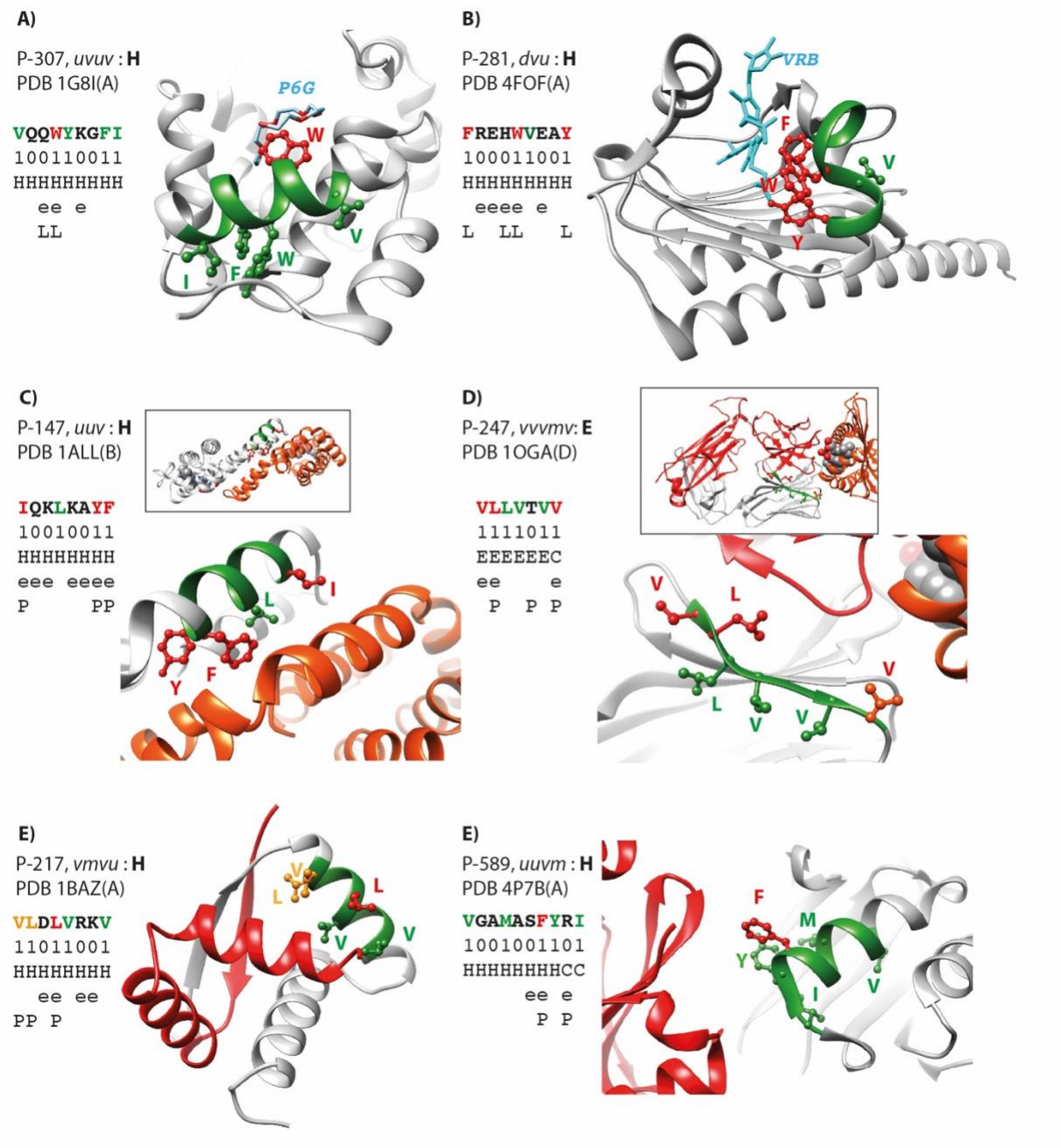


Figure 6: Concordant hydrophobic clusters with exposed hydrophobic amino acids. The position of the HC is highlighted in green on the ribbon representation of 3D structures. Only hydrophobic amino acids of the HCs are shown with exposed and buried amino acids depicted in red and green, respectively. The amino acid sequences of the HCs are shown, as well as their binary, Peitsch (P-) and Quark (Q-) codes, and the assigned secondary structure (H stands for "helix", E for "strand" and C for "coil"). Positions that are exposed to solvent (e = >36% relative accessibility) and reported in the PDBSUM as interacting with partners (P=protein, L=ligand) are also indicated. Non-exposed (buried) positions interacting with ligand (L) or other protein (P) subunit are depicted in orange. When the 3D structures are included in protein complexes, the other subunits are coloured orange/red on the 3D structure. **A)** Human frequenin (pdb 1G8I [23]) **B)** GAF domain of the blue/green photochromic cyanobacteriochrome (CBCR) from *Thermosynechococcus elongatus* (pdb 4GLQ [24]). **C)** Allophycocyanin from *Spirulina platensis* (pdb 1ALL [39]). The protein, which binds phycocyanobilin (CYC), is composed of two subunits, building up an (α and β) heterodimer, by convention termed monomer. **D)** T-cell repecor (pdb 1OGA [40]). **E)** Arc repressor from *Enterobacteria* phage p22 (pdb 1BAZ [41]): an example of

repressor proteins, which bind DNA through an antiparallel β-sheet formed by the two subunits of a dimer. **F)** *Salmonella typhimurium* peptidyl t-RNA hydrolase (pdb 4P7B [42]).

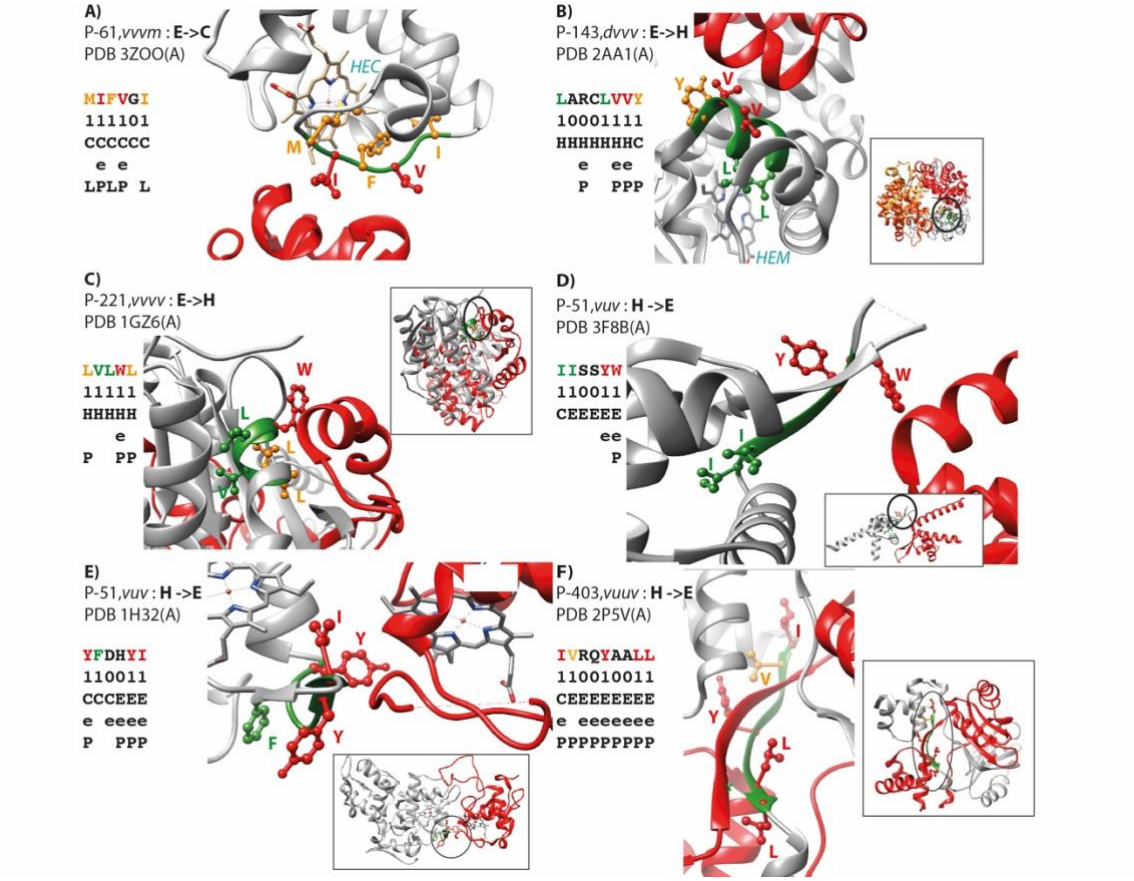


Figure 7: Discordant hydrophobic clusters with exposed hydrophobic amino acids. HCs and hydrophobic amino acids are depicted as described in Figure 6. A) Human cytochrome c (pdb 3ZOO [43]), B) *Trematomus newnesi* (3R)-hydroxyacyl-coA dehydrogenase (pdb 2AA1 [44]), **C)** *Rattus norvegicus* pyrroline-5-carboxylate reductase (pdb 1GZ6 [45]), **D)** *Lactococcus lactis* transcriptional regulator lmrr (pdb 3F8B [46]), **E)** *Rhodovulum sulfidophilum* reduced SoxAX complex (pdb 1H32 [47]), **F)** *Neisseria meningiditis* transcriptional regulator (pdb 2P5V [48]).

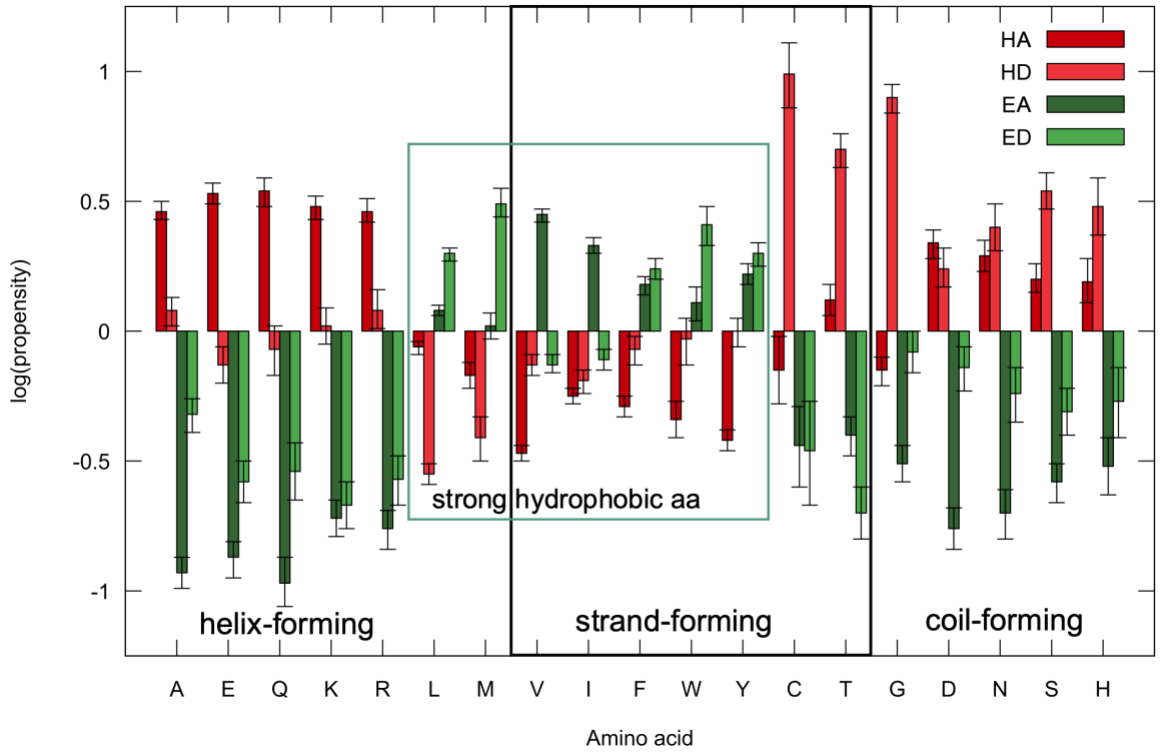


Figure 8: Differences of the amino acid propensities between the concordant (A) and discordant (D) states. Propensities are calculated as described in the Material and Methods section. They are reported for HC with α-helix (H - red) and β-strand (E - green) affinities. Amino acids (one-letter code) are grouped according to their overall propensities for α-helices, β-strands and coils [3].

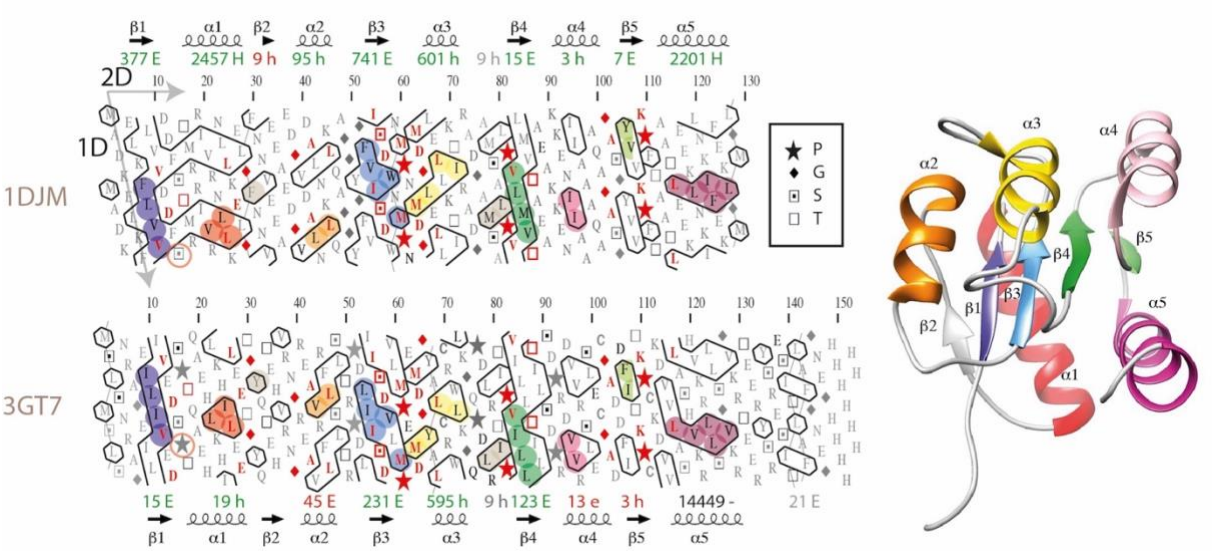


Figure 9: Fold signatures. HCs with clear affinities for the α (H) or β (E) states often correspond to core secondary structures and are much more conserved than the amino acid sequences. Therefore, they constitute fold signatures, which can be used to compare sequences sharing low levels of identity. This is exemplified here with two sequences of the same family, for which experimental 3D structures have been solved (pdb 1DJM and 3GT7). The way to interpret the 2D plot is as in **Figure 1**; observed secondary structures, as well as P-codes and secondary structure affinities of HCs are indicated on top and bottom. H stands for "helix", E for "strand" and C for "coil" with upper and lower cases corresponding to high and low affinities, respectively. Conserved hydrophobic amino acids between equivalent HCs are colored (with correspondence to the 1DJM 3D structure shown at right), corresponding to amino acids participating in the hydrophobic core (fold signatures), whereas sequence identities (21 %) are indicated in red. The first HC of the upper sequence can be split into two separate clusters at the level of serine 14, corresponding to proline 16 in the bottom sequence (encircled).