The respective roles of polar/nonpolar binary patterns and amino acid composition in protein regular secondary structures explored exhaustively using hydrophobic cluster analysis

Joseph Rebehmed, Flavien Quintus, Jean-Paul Mornon, Isabelle Callebaut

THE RESPECTIVE ROLES OF POLAR/NON POLAR BINARY PATTERNS AND AMINO ACID COMPOSITION IN PROTEIN REGULAR SECONDARY STRUCTURES EXPLORED EXHAUSIVELY USING HYDROPHOBIC CLUSTER ANALYSIS

JOSEPH REBEHMED, FLAVIEN QUINTUS, JEAN-PAUL MORNON, ISABELLE CALLEBAUT*

CNRS UMR7590, Sorbonne Universités, Université Pierre et Marie Curie-Paris6 – MNHN – IRD – IUC, Paris, France

*** Corresponding author:**

IMPMC, CNRS UMR7590, UPMC

Case 115, 4 place Jussieu

75252 Paris Cedex 05

France

e-mail: isabelle.callebaut@impmc.upmc.fr

Keywords: orphan sequences, globular domains, secondary structures, hydrophobicity, amino acid composition, concordance, discordance

Running title: Discordances from binary pattern preferences

ABSTRACT

Several studies have highlighted the leading role of the sequence periodicity of polar and nonpolar amino acids (binary patterns) in the formation of regular secondary structures (RSS). However, these were based on the analysis of only a few simple cases, with no direct mean to correlate binary patterns with the limits of RSS.

Here, we considered HCA-derived hydrophobic clusters (HC), which are conditioned binary patterns whose positions fit well those of RSS, and analyzed all the HC types, defined by unique binary patterns, which are commonly observed in 3D structures of globular domains. We observed that the 180 HC types with preferences for either α-helices or β-strands distinctly contain basic binary units typical of these RSS, a general trend thus supporting the "binary pattern preference" assumption. We also focused on HC for which observed RSS are in disagreement with their expected behavior (discordant HC). We distinguished HC types with moderate preferences for RSS, with "weak" binary patterns and versatile RSS and HC types with high preferences for RSS, with "strong" binary patterns and then displaying nonpolar amino acids at the protein surface. We show that in both cases, discordant HC can be distinguished from concordant ones by well-differentiated amino acid compositions.

The obtained results could thus help to complement the currently available methods for the accurate prediction of secondary structures in proteins from the only information of a single amino acid sequence. This can be especially useful for characterizing orphan sequences and for assisting protein engineering and design.

INTRODUCTION

The sequencing of the first chromosomes and genomes has highlighted a large number of genes, called "orphans", that had no detectable homologs to any other genes [1]. Even after the sequencing of many other complete genomes and the development of more sensitive homologous searches tools to detect more distantly related proteins, the number of orphan genes remained significantly high [2] and this is still true today, even though this notion is now refined by "phylostratigraphic" approaches, allowing to identify the phylogenetic level where the orphanicity is confined to [3]. At the protein domain level, many regions that are predicted to fold into globular domains do not match any known models stored in domain databases. The percentage of these orphan domains relative to the total number of globular domains ranges between less than 10 % in bacteria and more than 50 % in apicomplexa [4].

A critical issue when addressing the problem of orphans is to distinguish between proteins or protein domains that have rapidly evolved, those that have been lost in closely related species and those that were created *de novo* [3]. In the case of fast evolving proteins (or domains), detecting remote relationships remains a tricky task as the mean sequence identities are generally very low [5] and therefore, significant similarities can not been detected from background noise, even using the most sensitive programs.

A way to help the detection of remote relationships between sequences is to introduce structural information in the sequence comparison procedure. This information must not only relies on the position and nature of regular secondary structures (RSS: α-helices and β-strands), which are much more conserved than the sequence itself, but also on fold signatures, which involve conservation of hydrophobic features at specific positions [6]. The difficulty in the case of orphan proteins or protein domains is to gain this information from the

consideration of a single sequence, rather than from a set of homologous sequences, as usually made by current prediction tools.

These issues can be intrinsically addressed in a relatively simple way by considering the sequence periodicity of polar and nonpolar amino acids, which plays a critical role in the choice between α-helices and β-strands [7]. Hence, for β-strands and α-helices, this periodicity must place nonpolar amino acids every two, and three or four positions, respectively. The intrinsic preferences of amino acids for one secondary structure versus another can thus be overwhelmed by the drive to form amphiphilic structures capable of burying hydrophobic surface area and, as a consequence, the precise identity of a residue at a particular location in a sequence may be less important than the simple choice of whether it is polar or nonpolar. This binary pattern preference has also been inferred using a simple polymer model [8] and has been extensively used for designing *de novo* proteins [9]. Globally it helps to explain why a given fold can be encoded by many different amino acid sequences. However, it has been documented in only a few specific cases, especially for self-assembling oligomeric peptides [7]. Moreover, simple binary patterns have some limitations when applied to the analysis of secondary structures preferences, as noise can arise from the fact that they can include or they can be included in other binary patterns. For instance, as shown in **Fig. 1**, the binary pattern 11 (1 representing an hydrophobic amino acid and 0 any other amino acid) can be inserted in many binary patterns (three of which are shown on the figure) associated with different secondary structures. As a consequence, there is no simple way to define the pattern that matches at best the secondary structure limits.

A refinement of the binary pattern concept was introduced with the Hydrophobic Cluster Analysis (HCA) methodology [10,11], which gives access to *conditioned* binary patterns, called hydrophobic clusters (HC), through a 2D transposition of the protein sequence shown in

Fig. 2. The definition of HC is indeed conditioned by the presence of a minimal number of nonhydrophobic residues (called *connectivity distance*) or a proline residue that separates two hydrophobic amino acids belonging to two distinct HC (**Figs. 1 and 2**). The connectivity distance of four (that of an α-helical net) and the alphabet including seven strong hydrophobic amino acids (V, I, L, F, M, Y, W) have been demonstrated to provide the best correspondence between HC and RSS [12]. Following this definition, conditional binary patterns can not be intertwined, *i.e.* they can not include or be included in other binary patterns. Therefore and importantly, they carry a much more differentiated information and are much more correlated to the RSS limits than non-constrained binary patterns [13] (**Fig. 1**). In line with these methodological aspects, the HCA approach has been successfully used to overcome limitations of sequence similarities searches in detecting remote relationships, by adding this relevant 2D information, which can be rapidly deduced from a single sequence. Indeed, as HC constituting the structural core of domains are much more conserved than the sequence itself, they can be used to highlight conserved fold signatures, far beyond a simple three-state secondary structure classification (α, β or coil) and despite very low levels of sequence identities [14-18]. Tools have also been recently developed to predict automatically the limits of regions with high density in HC ("foldable" domains) [4], and have been combined with domain database searches for detecting orphans within whole proteomes [19] and revealing hidden relationships [20-22].

In this study, we thus wished to analyze the 294 most frequent HC types (each type being defined by a unique binary pattern), which are commonly encountered in known 3D structures of globular domains and are stored in a dictionary established in a previous work [23]. These HC gather more than 80 % of the total number of observed clusters. We considered here the HC types mainly associated with α-helices (97 HC types) and with β-strands (83 HC types),

each category can be further classified into HC with strong and moderate propensities for these SSR. We discarded HC mainly associated with coils (binary patterns 1 and 11) and those that are called multiple (HC associated with at least two different RSS). **Figure 2** provides two examples of HC types with strong propensities for α-helices (HC associated with helix D, in orange) and for β-strands (HC associated with strand β1, in blue). This protein sequence also contain a large "multiple" HC, including three SSR (helices αB (purple) and αC (pink) and strand β2 (medium blue)) separated by short loops. Finally, it also highlights two HC with low propensities for β-strands, one of which indeed corresponding to a β-strand (strand β3, in light blue) while the other is in this particular case associated with an α-helix (αA, in red).

In this study, we first analyzed, in an exhaustive way, the relationship between the periodicity of polar/nonpolar amino acids of these common HC and their preferences towards one of the secondary structure states. To our knowledge, most of the previous studies were limited to only a few typical binary patterns and such a comprehensive investigation has never been made for the ensemble of conditional binary patterns [7]. This is however an important issue for definitely assessing the preponderant influence of binary patterns over amino acid preference on secondary structure formation, with the additional benefit here that a better correlation is observed between binary pattern and SSR through HC. To that aim, we used a specific binary code, the Quark (Q-) code, to decompose each HC into its basic units. Indeed, each HC can be unequivocally described as a combination of four basic Q-codes, which are defined along the three axes of the helicoïdal representation: (i) V (vertical) =11 and M (mosaic) =101, (ii) U (up) = 1001 and (iii) D (down) = 10001 (**Fig. 2**). Hence, the binary pattern corresponding to strand β1 (1010101) can be translated in MMM, that of helix αD (10011001) in UVU. As M and U/D correspond to the non-polar amino acids periodicity observed in β-strands and α-

helices, respectively, considering a HC as a Q-code combination allows evaluating how its main periodicity correlates with its main state in terms of RSS. In this way, we were be able to confirm, at large scale, the binary pattern preference, *i.e.* that the secondary structure predominantly observed (major or concordant states) well corresponds to that expected from the conditioned binary pattern (**Fig. 3a**).

We also addressed here the second issue of HC that do not adopt the secondary structures expected from their binary pattern (minor or discordant states). Those situations are not rare, especially for HC with medium propensities for α-helices (h) and β-strands (e). This is exemplified with the HC associated with α-helix A in **Fig. 2**, which corresponds to an HC type with moderate affinity for β-strands. In this specific case, such discordance may be explained by the fact that the binary pattern is not strong enough to drive by itself the formation of a particular SSR. In contrast, in the case of clusters with strong propensities for α-helices (H) and β-strands (E) and which have binary patterns very typical of the preferred SSR, such a discordance may result in the exposure of non-polar amino acids to solvent, which may participate in interaction sites (**Fig.3b**). We showed here that in both cases, these discordant clusters can be distinguished from concordant ones on the basis of their composition in amino acids, thereby allowing to define situations where the amino acid composition takes precedence over the binary pattern.

Altogether, the results presented here, based on the original concept of HC and deduced from a comprehensive analysis of 3D structures databases, bring new information to clarify the roles played respectively by binary patterns and amino acids in the formation of regular secondary structures. They provide insightful information to predict secondary structures in

proteins from the only information of a single amino acid sequence, which can be used to decipher orphan proteins or protein domains.

MATERIAL AND METHODS

DATABASES We first extracted hydrophobic clusters (HC) from a representative set of globular domains 3D structures. To that aim, we considered the ASTRAL SCOP database (version 2.03, 2014-01-22) of protein structures [24] at the sequence identity thresholds of 95% and 40% to treat the redundancy at two different levels. We then selected protein structures that belong to the first five SCOP classes (a, b, c, d and e) and that were solved by X-ray diffraction with a resolution lower than 2.5 Å. We discarded chains of the first five classes (a, b, c, d and e) that are also reported in other SCOP classes. All NMR structures were excluded from the analysis, as well as structures that belong to SCOP classes f (Membrane and cell surface proteins and peptides) and g (Small proteins). We also eliminated models and files containing only Cα coordinates or missing a lot of residues. 15245 and 8507 entries (out of 22935 and 12198 entries in SCOP 95 and SCOP 40) were thus carefully chosen, respectively, for further analyses.

SECONDARY STRUCTURE ASSIGNMENT Secondary structures were assigned using DSSP program [25] (H-bond energy calculations), starting from the atomic coordinates of the selected 3D structures, extracted from the Protein Data Bank. The assigned secondary structures were then grouped into 3 categories: **H** for α-helix (H), 3_{10}-Helix (G) and π-Helix (I); **S** for β-bridge (B) and extended strand (E); **C** for hydrogen bonded turn (T), bend (S) and loop or irregular (blank).

HYDROPHOBIC CLUSTER DEFINITION The next step consisted in extracting hydrophobic clusters

(HC) from both 3D structure datasets. As described in the introduction, an HC is defined as a succession of strong hydrophobic residues separated from other HC by breakers that are composed of at least four consecutive non-hydrophobic amino acids (*connectivity distance*) or a proline (**Figure 2**). The strong hydrophobic residues are V, I, L, M, F, Y, W in the standard use of HCA approach, in which the α-helix is used as a two-dimensional support for the 2D HCA transposition of the sequence [10,11]. These parameters have proven to be optimal, providing the best correspondence between HC and RSS [12].

HC types are currently designated using their binary codes, where 1 stands for any strong hydrophobic amino acids and 0 stands for the other amino acids, except for proline. As the smallest HC, containing only one (1) or two hydrophobic amino acids (11), are not frequently associated with RSS [23], these clusters were excluded from the analyses.

To each binary code corresponds a Peitsch (P)-code that 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 of the HC corresponding to 0). Hence, for the asymmetric HC with binary code 1101, the P-code is $1x2^3 + 1x2^2 + 0x2^1 + 1x2^0 = 13$. A cluster code converter (binary to Peitsch and Peitsch to binary) can be found at http://osbornite.impmc.upmc.fr/hca/converter/index.html.

Finally, we also examined here the different HC types in the light of the combination of Quark codes that allow us to consider the periodicity of hydrophobic positions. Quark codes consist in the basic cluster units following the three axes of the HCA two-dimensional representation: (i) V (vertical) = 11 and M (mosaic) = 101, (ii) U (up) =1001 and (iii) D (down) = 10001 (**Figure 2**).

CONCORDANCE/DISCORDANCE We studied whether the HC are concordant or discordant in regards to their secondary structure affinity that was previously determined in the dictionary of HC [23] (http://osbornite.impmc.upmc.fr/hca/HCA-table.html). This dictionary lists the 294

most frequent clusters types (which represent over 80% of the total number of clusters of all types, at the time of its conception), with their associated frequency, propensity and secondary structures affinity (α helices, β strands or "coil"). Among them, 97 and 83 species appear predominantly associated with α helix and β-strand structures, respectively, after their propensities being calculated for the different states. HC can be further divided in HC with high affinities for a given secondary structure (upper cases H and E) if other propensities are lower than 1 or the difference between the highest propensity and the second one is greater than 1, and HC with low affinities (lower cases h and e) otherwise.

We considered here various degrees of strictness for the definition of the concordance/discordance states: HC were defined as concordant if at least 80 % (80-level) or 60 % (60-level) of their positions are observed in the HC major state (as reported in the dictionary), as discordant if less than 20 % are observed in this same HC major state, and as intermediate otherwise (**Supplementary Data 3**).

AMINO ACID PROPENSITIES We quantified the intrinsic preferences of amino acids for the different RSS in the whole datasets and in the concordant/discordant states by calculating their propensities according to the following equation:

$$P_{i,ss} = \frac{N_{i,ss}/N_{ss}}{N_i/N_{Tot}}$$

where $N_{i,SS}$ is the number of amino acid *i* in a specific secondary structure (Helix, Sheet or Coil), N_{SS} is the total number of amino acids found in this specific secondary structure, N_i the number of amino acid *i* found in all three structures categories and N_{Tot} is the total number of amino acids.

A value of propensity greater than one (or positive if logarithms are used) indicates a higher

preference of the amino acid in this regular secondary structure whereas a value less than one (or negative if logarithms are used) implies lower preference of that amino acid. These calculations were made on the entire protein sequences and were also restricted to the hydrophobic clusters in general and more specifically to the RSS concordant and discordant states.

Comparison between amino acid occurrences in concordant and discordant HC were analyzed by Pearson's χ^2 test.

DATA VISUALIZATION Structures were analyzed and manipulated using the PyMOL Molecular Graphics System (version 1.7, Schrödinger, LLC, USA). The sequence logos of the amino acids and secondary structures were generated using WebLogo version 3.3 [26].

RESULTS AND DISCUSSION

REFERENCE DATASETS

3D structure database

Two sets of 3D structures were derived from the SCOP database (classes a, b, c, d, e), considered at two levels of sequence redundancy (ASTRAL 95 (less than 95 % sequence identity between sequences) and ASTRAL 40 (less than 40 %)). We considered these two levels of redundancy (SCOP 40 and SCOP95) in order to have access to a sufficient number of data for each hydrophobic cluster (HC) type (a type being defined by a unique binary code, see below). We however ensured that results, presented below for the SCOP95 database, are similar between the two datasets and that no obvious bias could be detected from their comparison. 15245 and 8507 entries (total number of amino acids of 2,873,086 and 1,618,101 respectively) were kept after filtering these sets using several criteria (see Material and Methods). In the remaining text, the final sets are referred to as the SCOP95 and SCOP40 databases. The amino acid compositions of these two databases are given in **Supplementary Data 1**.

Hydrophobic Clusters

We extracted hydrophobic clusters (HC) from these SCOP95 and SCOP40 databases and classified them into HC types according to their binary codes. To each binary code corresponds a numerical Peitsch (P-) code. This gives a meaningful way of HC type classification (**Fig. 2 and Material and methods**). We retained for further analyses the 294 HC types which are commonly observed in globular domains, as defined in our previously published dictionary [23]. The two smallest types (binary codes 1 and 11) were however not considered further as they are mainly associated with coils [23]. The total numbers of HC are 155712 (SCOP95) and 87876 (SCOP40), of which 113960 (73.2 %) and 64032 (72.9 %) are

associated with the 292 common HC types (**Supplementary Data 2A**). The selected HC thus represent the large part of the total number of HC present in the recent SCOP databases. We focused our study on the HC types with propensities for α-helices (97 HC types) and with β-strands (83 HC types), as defined in our previously published dictionary [23], and discarded "multiple" HC (HC associated with at least two different RSS – 112 types). In the SCOP95 (*SCOP40*) database, the HC types with affinities for α-helices and β-strands include a total of 40434 (*23076*) and 61698 (*34390*) HC, respectively (**Supplementary Data 2A**).

A further and important distinction can be introduced between HC types with strong (upper cases) and moderate (lower cases) propensities for α-helices (H (48 types) and h (49 types), respectively) and β-strands (E (41 types) and e (42 types)), respectively). This distinction was made in our previously published dictionary[23], in which HC were assigned as having high affinities for a given secondary structure (H and E) if propensities for other secondary structures are lower than 1 or the difference between the highest propensity and the second one is greater than 1, and as having low affinities (h and e) otherwise.

On average, each HC type of the SCOP95 (*SCOP40)* database with propensities for α-helices and β-strands contains 417 *(238)* and 743 *(414)* HC, respectively (**Supplementary Data 2B)**. However, as indicated by high standard deviations (α-helices HC: 926 *(532)*, β-strands HC: 1679 *(920)*), there are large variations of HC occurrences between HC types, the smallest ones being the most populated (**Supplementary Data 2D)**. For instance, there are 11823 HC with P-code 5 (101, length 3) and only 37 HC with P-code 1637 (11001100101, length 10) in the SCOP95 database.

Lengths of HC types considered here vary between 3 and 13 amino acids (**Supplementary Data 2C**), the mean length of HC types associated with α-helices and β-strands being 7.9 (e), 6.6 (E), 9.2 (h) and 9.4 (H). The database thus include HC types matching nearly all of the β-

strands and a large part of the α-helices [23]. HC types corresponding to some large α-helices are missing, as they are not sufficiently represented at the time of the dictionary construction. As small HC are more abundant than large HC, the mean lengths of the total number of HC, in the SCOP95 (*SCOP40)*, associated with α-helices (h:6.2 *(6.2)*, H:6.9 *(6.7)*) and β-strands (e:4.6 (*5.0*), E:4.8 (*4.8*)) are smaller than the corresponding mean length of HC types (**Supplementary Data 2D)**.

Here, it should also be pointed out that, albeit having a unique binary pattern, each HC type covers a wide variety of sequences. Thus, pairwise sequence identities calculated by HC type show comparable distributions for HC derived from the SCOP40 and SCOP95 databases (mean sequence identities of 15.62% and 15.45% respectively) (**Fig. 4-A**). As shown in **Fig. 4-B** for the two HC types shown in **Fig. 2,** strongly associated with α-helices (P-code P-153) and β-strands (P-code P-85), it is not rare that two HC of the same type share 0 % sequence identities, adopting either similar or opposite secondary structures.

Finally, we analyzed the HC types in light of the observed SSR and for each of them, we separated HC for which the observed regular secondary structures (SSR) are in agreement (concordant HC) and in disagreement (discordant HC) with the SSR expected from the dictionary[23]. We fixed here two levels for the definition of the concordance state, where a concordant state is assigned if at least 80 % (80-level) or 60 % (60-level) of the HC positions are observed in the major state (as reported in the dictionary) and as discordant if less than 20 % are observed in the same major state, and intermediate otherwise *(see Material and Methods and **Supplementary Data 3A**)*. These definitions are fairly stricter than the rule previously used in our dictionary[23], where only one position assigned in a given SSR was sufficient to assign the HC in this state. At that time, this last assignment rule was however supported by a general large coverage of HC types by SSR. Here and logically, less

concordant HC are observed for the strictest 80-level than for other more permissive levels (e.g. 60 %) (**Supplementary Data 3B**). When the definition rule of concordant states is less strict, the number of intermediate states decreases at the benefit of the concordant states, while the discordant state remains stable. However, this strictest level (80 %) enables us to have access to the most unbiased information in order to better distinguish between the two states based on the amino acid composition.

The number of concordant and discordant HC is variable depending on the HC type class (H, h, E, e) and within a class, on the HC type itself. As shown **Supplementary Data 3B**, the mean occurrence of concordant HC is greater than that of discordant HC for HC types with high propensities for SSR, whereas these values are roughly similar for HC types with moderate propensities for SSR. However, when HC types are considered individually and especially when the concordance level is fixed to 60%, most of them are more concordant than discordant HC (**Supplementary data 3C and 3D**). Indeed, the means of the ratios between the number of concordant and discordant HC per HC type (Rcd) are greater than 1 whether HC types with high or small propensities for α-helices and β-strands. This supports the general preferences previously reported in our dictionary [23]. Values are logically higher for HC types with high propensities for α-helices and β-strands, with maximal values obtained for the former. In our databases, some HC with moderate affinities for SSR, especially belonging to the e class, are however more discordant than concordant, adopting more frequently opposite SSR to that expected from the dictionary (**Supplementary data 3D**).

The 2D HCA plots of all the HC types with moderate and high affinities for α-helices (h, H) and β-strands (e, E) are reported in **Supplementary Data 4**, classified as a function of the increasing Rcd ratio. They clearly emphasize the marked 2D characteristics of HC that are frequently associated with either α-helices or β-strands (HC with high affinities for α-helices

(H) and β-strands (E) and HC with moderate affinities but with high Rcd ratio). These will be discussed below.

PREFERENCES OF HC BINARY PATTERNS FOR RSS

In order to evaluate if the binary pattern information can be significantly linked to the SSR information, we examined the RSS affinities of HC types, in the light of their sequence periodicities of polar and nonpolar amino acids.

We first addressed this issue by analyzing the two representative HC types with P-codes 153 (10011001) and 85 (1010101), depicted in **Fig. 2**. While they possess the same number (4) of hydrophobic residues, they are highly associated with α-helices and β-strands, respectively. HC types of P-codes 153 (*697 and 396 occurrences in SCOP95 and SCOP40, respectively*) and 85 (*511 and 297 occurrences in SCOP95 and SCOP40, respectively*) are indeed mainly found, in our previous dictionary [23], associated with α-helices (81 %) and β-strands (70 %), respectively.

In the very strict definition of the concordance state (80-level) we adopted here (see above), these HC still conserve a high rate of association with α-helices (67 %) and β-strands (44 %), respectively (**Supplementary Data 5**), while for more permissive definition (60-level), the association rates tend to those previously calculated (82 % α-helices for P-153 and 55 % β-strands for P-85 – **Supplementary Data 5**) [23]. The WebLogo representations shown in **Fig. 4** give a meaningful view of the composition in amino acids (last two columns) and the corresponding observed secondary structures (first column) per position of the HC. They clearly indicated the strong preferences of the species P-153 and P-85 for two distinct RSS (α-helices and β-strands) over the whole length of the HC. The left panel (DSSP secondary structure assignments) clearly indicates that the HC limits well cover the SSR limits for the concordant states (observed SSR in agreement with that expected from the dictionary),

whereas some shifts are observed for the discordant states (observed SSR in disagreement with that expected from the dictionary). A similar behavior can be found for most of the HC considered here (Weblogos of the whole set of HC can be found in **Supplementary Data 6**). The information to withhold from these examples is that in the HC species with P-code 153 (10011001), nonpolar amino acids are found every three or four positions, whereas nonpolar amino acids are observed every two positions in the HC species with P-code 85 (1010101). The major SSR state observed for these two HC species is thus in accordance with the binary pattern.

We then quantified, in a comprehensive way, the general trends of HC with hydrophobic amino acid periodicities corresponding to α-helices and β-strands, respectively, to effectively adopt such RSS. To that aim, we used Quark codes (Q-codes) to decompose each HC into its basic units. According to the connectivity distance of 4, which is commonly used in the HCA approach, there are only four Q-codes along the three axes of the helicoïdal representation: (i) V (vertical) =11 and M (mosaic) =101, (ii) U (up) = 1001 and (iii) D (down) = 10001, whose combinations can be used to describe any HC (**Fig. 2**). Hence, the P-153 HC type (binary code 10011001) can be decomposed into 1001 (U) + 11 (V) + 1001 (U), thus UVU and the P-85 HC type (binary code 1010101) into 101 (M) + 101 (M) + 101 (M), thus MMM. According to the binary pattern preference, hydrophobic amino acid periodicities associated with helices (one hydrophobic amino acid every three or four positions) should mostly correspond to the U and D Q-codes (horizontal shapes of the clusters), whereas those associated with strands (one hydrophobic amino acid every two positions) should be rather observed with the M Q-codes (vertical shapes of the clusters). The V-code (made of two consecutive hydrophobic amino acid) can be observed either in α–helices (hydrophobic face of an amphipathic helix) or in β-strands (buried beta strands). That is what is actually

observed for the P-153 HC (H affinity, Q-code UVU) and P-85 HC (E affinity, Q-code MMM).

When broadening the analysis to the whole set of HC (**Fig. 6**), we observe that HC with helix affinity (H and h, in blue) are mainly composed of combinations of D (15 and 18 %), U (37 and 23 %) and V (39 and 38 %) Q-codes, whereas HC with strand affinity (E and e, in green) principally include M (37 and 31 %) and V (57 and 49 %) Q-codes. The D Q-codes are even absent in E clusters, whereas the M Q-code is present at only 9 % in H clusters. In this last case, the tolerance may be due to the ability of α-helices, longer than β-strands, to locally accommodate this small binary pattern deviance, as well as to the relative independence of α-helices relative to non-local interactions (internal network of H-bonds). It is also possible that these account for the few HC that actually do not form the expected SSR.

On another hand, there is no clear correlation between the length of the Q-code and specificity of the secondary structure, as the smallest one, the V Q-code, is found at similar rates in both α-helices and β-strands (45 % and 55 %, respectively, **Fig. 6, top panel**). This is also supported by the analysis of the relative occurrence of Q-codes as a function of the HC type length (**Fig. 6, bottom panel**), which clearly shows that the smallest Q-codes, the V Q-code, but also the M Q-code (specific of β-strands), are also found in HC with large length.

In conclusion, these first results demonstrate, in an exhaustive way, that the periodicities in polar and nonpolar amino acids indeed play a predominant role in the formation of a particular RSS. This is particularly true for HC with strong propensities for the α (H) and β (E) states, in which D/U and M Q-codes are predominant and which are much more observed in a concordant state than in a discordant one. These characteristics can be visually recognized on the HCA plots (**Supplementary Data 4**), as these HC types have typical horizontal and vertical shapes, respectively. HC with moderate propensities for the α (h) and β (e) states exhibit more "mixed" Q-codes combinations (**Fig. 6**) and their 2D shapes are

generally not meaningful (**Supplementary Data 4**). This suggests that for these HC (especially those for which similar levels of concordant and discordant HC are observed, see below), the binary patterns should not be enough informative and that the amino acid composition may play a critical role in determining which kind of SSR is adopted.

AMINO ACID SEQUENCES OBEYING AND DISOBEYING THE BINARY PATTERN PREFERENCES.

When the binary pattern (topological information) is strong enough to drive the formation of SSR (H and E HC, with more concordant than discordant occurrences), the observation of a SRR opposite to the expected one (β-strand for H HC and α-helix for E HC) means that the HC is likely to expose some hydrophobic positions to the solvent, as depicted in **Fig. 3B**. Detecting such clusters is of interest for predicting interaction sites.

However, for some HC, a similar or higher number of discordant cases are observed (**Supplementary Data 3 and 4**). As mentioned before, these have principally weak affinities for the α- (h) or β- (e) states and are mainly composed of mixed α/β Q-codes or Q-codes opposite to their previously defined affinities. For instance, P-277 (h, DMM, 100010101) has a mixed behavior (with propensity for the α-state in its N-terminal part and for the β-state in its C-terminal part). This is also the case of P-89 (e, MVU, 1011001), shown in **Fig. 2**, which forms in this example a α-helix. In these particular cases (h and e HC), the binary pattern is not enough informative to drive the formation of SSR.

We hypothesized that in both cases (HC types with high and moderate propensities for SSR) and beyond the consideration of the binary pattern information, the amino acid composition may be critical for distinguishing between concordant and discordant behaviors.

In regard to our archetypical HC types with P-codes P-153 (α-helix HC) and P-85 (β-strand HC) (**Fig. 5**), amino acid sequence profiles clearly differed between the concordant and

discordant states. For instance, for P-153 in the discordant state, hydrophobic amino acids with strand propensities (V, I, F, Y) outperforms those with helix propensities (L, M), whereas the level of A, E, Q, K, R also decreases. Loop-forming residues markedly punctuate the beginning and the end of the cluster. In the P-85 discordant state, L and A take priority over V/I and T in hydrophobic and non-hydrophobic positions, respectively.

We thus analyzed, in a systematic way, the amino acid composition within HC predominantly associated either with α-helices or with β-strands and then distinguished concordant and discordant behaviors (*i.e.* with observed secondary structures corresponding or not to that expected from the dictionary, respectively).

To that aim, we first calculated, as a reference, the frequencies and associated propensities of amino acids for the different secondary structure states (**Fig. 7 A-B** (radial graphs) and **Supplementary Data 7** (bar plot of logarithms)) in the proteins extracted from SCOP95 and SCOP40. The deduced propensities are similar whatever the redundancy level is (**Fig.7-A** (SCOP95) and **7-B** (SCOP40)) and are consistent with those reported in previous works [11]. Briefly, these propensities highlight three classes of amino acids with preferences for α-helices (A,L,M,E,Q,K,R - blue), β-strands (V,I,F,W,Y,C,T - red) and coils (P,G,D,N,S - yellow). Histidine (H) has quasi-equal preferences for the three states. Regarding the aliphatic hydrophobic amino acids (M, L, I, V), the propensities of M and L are almost similar for the α and β states, whereas propensities of V and I for the β state are significantly higher than for the α state. The same calculations were made at the level of HC, rather than on the whole protein sequences (**Fig. 7-C**). As expected from the fact that HC are mainly associated with RSS, loop-forming residues (G,D,N,S) are less frequent within the limits of HC (data not shown). Note that proline (P) is excluded from these calculations, as it does not participate in

HC. However, when present, loop-forming residues are associated with higher propensities towards coils, highlighting their preferences for coil regions included within the limits of HC (N- and C- terminal limits). Other residues share similar trends towards RSS than when calculated on the whole protein sequences.

We then calculated propensities of each amino acid for the concordant and discordant states at the level of the entire dataset, within the HC limits (without considering amino acids before and after the HC N- and C-terminal hydrophobic amino acids). Comparison of occurrences and propensities between the concordant and discordant states **(Fig. 7-D** to **7-F**) indicates a clear difference of behavior between HC typical of α-helices and β-strands. Significant χ^2 (values (>28.89, $p<0.05$ and >42.81, $p<0.001$) are observed between the concordant and discordant distributions, with per amino acid χ^2 values below the χ^2 α value only for I (χ^2=12.33) and D (χ^2=2.05) in the case of HC typical of α-helices and F (χ^2=6.57), W (χ^2=17.30), Y (χ^2=2.45) and H (χ^2=4.91) in the case of HC typical of β-strands. For α-helix forming HC, a clear increase of the propensities in discordant states is observed for the loop-forming glycine and for threonine and cysteine (often associated with β-strands) at the expense of all α-forming amino acids (alanine, methionine, leucine, as well as glutamic acid, glutamine, lysine and arginine) (see in **Fig. 7-D** the shift of the blue points at left towards the green ones at right). Propensities of the aromatic β-strand forming residues W and Y also increased. Thus, for HC typical of α-helices, the overall load of α-forming residues, both in hydrophobic and non-hydrophobic positions, is diminished in the discordant state for residues favoring β-strands (especially C and T, but also W and Y) and coils (especially G). The reason behind the fact that non-hydrophobic positions play a significant role in addition to hydrophobic ones is likely that these are generally involved, in α-helices, in intra-helical side chain-side chain interactions, contributing to the overall stability of the local structure [35,36].

Mimetic residues, such as C and T, which can be buried in a hydrophobic environment, play in this case a critical role for evolving non-hydrophobic positions towards hydrophobic-like ones. Integrating them in the hydrophobic alphabet indeed leads to the evolution of the HC towards binary patterns typical of β-strands.

For β-strand-forming HC, the situation appears different, as hydrophobic residues seem to play a more prominent role. Discordant propensities increase for the α-forming leucine and methionine, whereas those for valine and isoleucine (the β-strand-forming aliphatic amino acids) largely decrease (see in **Fig. 7-D** the shift of the red points at the bottom towards the yellow ones at left). Cysteine and threonine (which may behave as hydrophobic amino acids) also show a slight decrease. No clear impact is observed with the aromatic residues (tryptophan, tyrosine, phenylalanine), whilst these have high propensities for both the α and the β states. The reason behind the fact that hydrophobic amino acids are particularly affected in discordant strand-forming HC is probably that these are involved in the stabilization of this secondary structure through non-local interactions within the protein domain, while non-hydrophobic amino acids play a less prominent role. The critical role of hydrophobic amino acids in HC types with propensities for β-strands can also be well appreciated in **Supplementary Data 7B (bottom panel)**, where the couples of hydrophobic amino acid propensities for the concordant/discordant states clearly map apart from those of other amino acids. All these trends are also illustrated in **Fig. 7-E and 7-F** (for the helix- and strand-forming HC), with more pronounced propensities when discordances are defined very strictly (Disc80, with more than 80 % of the position adopting the regular secondary structure opposite to the major one).

In order to take into account the particular behavior of the few HC mentioned above (with mixed concordant/discordant behaviors), the propensities were also calculated by

distinguishing HC with high (H, E) and weak (h, e) affinities for the α and β states (**Supplementary Data 8 – top panel**). Interestingly, differences of the amino acid distributions between H and h HC (either in the concordant or discordant states, left panel) are significant but very low (χ^2 values of 119 and 100 (>28.87, p<0.05)), but with none of the per amino acid χ^2 values below the χ^2 α value). In contrast, striking differences are observed for HC with low (e) and high (E) strand affinities, either in the concordant or discordant states (right panel), as assessed by significant χ^2 values (>28.87, p<0.05) for the whole set of amino acids, with per amino acid χ^2 values below the χ^2 α value only for His (χ^2=21.87), M (χ^2=24.82), W (χ^2=0.91) and N (χ^2=19.28) in the concordant state and for V (χ^2=21.50), I (χ^2=6.15), C (χ^2=6.47) and G (χ^2=0.26) in the discordant state. For instance, T and C are more favored in e-HC than the strong aliphatic residues V and I, in opposition to the E-HC behavior. This suggests that the hydrophobicity of HC with low affinities for the β-state is less pronounced, allowing them to be more versatile in terms of SSR propensities.

This is also visible in the bottom panel of **Supplementary Data 8**, representing amino acid propensities of HC with high and low SSR affinities, analyzed in light of the secondary structure which is actually observed (helix, strand and coil). Again, no clear difference can be observed for HC with either strong (H) or weak (h) affinities for the α state: propensities of helix-forming amino acids (blue) globally decrease at the benefit of strand-forming ones (red -especially C and T, but also hydrophobic ones) for the strand assignment and of coil-forming ones (green- especially G) for the coil assignment. In contrast, a clear difference can be noticed for HC with strong (E) and weak (e) affinities for the β state. The clear predominant role of aliphatic hydrophobic amino acids (M, L versus V, I) is less pronounced for HC with weak β affinities at the benefit of non-hydrophobic amino acids. This clearly highlights the differentiated sensitivity of binary patterns typical of α-helices and β-strands to amino acid composition.

CONCLUSION

Some earlier studies have analyzed the respective importance of amino acid composition and polar/non polar binary patterns in determining the fold of globular proteins. On the one hand, pioneer studies, describing amino acid propensities for secondary structures based on a limited set of experimental 3D structures, have been followed by a large number of analyses, calculating global but also position-specific amino acid propensities for RSS (see [27] for a review). On the other hand, several studies have also shown that binary patterns are major determinants of the types of secondary structures [7], in addition to the fact that they generally deviate from randomness [28-32]. In this respect, binary patterns typical of α-helices are more favored than those typical of β-strands, an observation which was suggested to correlate with the necessity of avoiding aggregation of partially folded intermediates during intracellular folding. Hence, selection should have occurred to avoid the inherent aggregation tendency of β-strands [33,34]. This general behavior relative to randomness was also supported through the means of the conditioned binary patterns derived from HCA [13], which were considered in the present study.

However, only few studies focusing on few binary patterns have approached the combined influences of amino acid composition and binary patterns by testing the ability of the latter ones to modulate the amino acids secondary structure propensities. Hence, on an experimental level, it was shown for self-assembling oligomeric peptides, that the polar/nonpolar periodicity overwhelms the intrinsic propensities of amino acids [7]. This conclusion was based on the observed structural behavior of a few peptides, which were designed by considering binary patterns consistent with or opposite to amino acid intrinsic propensities towards secondary structures. A similar conclusion was drawn on a theoretical level with a simple

protein coarse-grained model, containing three types of residues (hydrophobic, polar and neutral) [8]. This one has been used to test the ability of binary patterns to tolerate opposite secondary structure propensities, which were introduced in the simulation by tuning the dihedral potential term in the force field. The originality of the study presented here was to focus on the whole set of binary patterns actually encountered in 3D structures of globular domains, instead of focusing on a few binary patterns tested in an artificial way for their ability to tolerate amino acids with unfavorable secondary structure propensities. The additional originality of this study is the use of HCA conditioned binary patterns (hydrophobic clusters or HC), which match well the positions of regular secondary structures [12] and thus carry out a much more differentiated information than simple binary patterns [13]. We focused here on the most frequent HC, which correspond to α-helices and β-strands commonly observed in globular domains, as deduced from our previously established HC dictionary [23].

We first demonstrated that HC types mainly associated with α-helices and β-strands, deconstructed into basic binary units (Q-codes), indeed contain binary patterns typical of α-helices (majority of D and U Q-codes) and β-strands (M Q-codes), respectively. This is in agreement with earlier studies focusing on a small number of binary patterns [29]. This is particularly true for HC types with strong propensities for α-helices and β-strands, while for HC types with moderate propensities (especially for β-strands (e)), more mitigated or ambiguous Q-codes are observed, often combining both binary signatures. Indeed, ambiguous helix-forming HC often includes M Q-codes (e.g. P-code 413 (110010101, VUMM, h)), whereas U and D Q-codes are observed in ambiguous strand-forming HC (e.g. P-code 81 (1010001, MD, e)). As a consequence, as observed on WebLogos (**Supplementary Data 6**), the global structure assignment (combining concordant and discordant HC) is generally not

uniform, following the binary pattern preference. A special attention should thus be given to these particular few clusters with ambiguous behaviors and for which the binary pattern is not sufficient to drive the formation of one particular kind of SSR. However, as suggested here, the amino acid composition should be sufficiently discriminant to enhance the prediction of the secondary structure actually adopted by these ambiguous HC. Consistent with this ambiguous behavior, the amino acid content of HC with moderate propensities for β-strands (e) is clearly distinct from that of HC with high E propensities, favoring amino acids that are not included in the HCA hydrophobic alphabet (*e.g.* threonine or cysteine), but behave as such depending on the environment. In these particular cases, integrating these amino acids in the hydrophobic alphabet may lead to evolve towards another binary code, meeting the RSS binary pattern preference. It is also possible that the weak level of strong hydrophobic amino acids encountered in these particular HC may be associated with a particular conformational plasticity. Including NMR 3D structures in our dataset should help to investigate such a particular issue.

Analysis of amino acid composition may also orientate the prediction of discordance states for HC having strong propensities for either α-helices or β-strands. This has clear topological implication as in this case, some of the hydrophobic amino acids belonging to the binary pattern are likely exposed to solvent (**see Fig. 3B**). Hence, hydrophobic clusters as defined by the HCA approach may provide interesting methodological tools for distinguishing hydrophobic amino acids involved in the protein core from those which are implied in the interaction with partners (proteins, nucleic acids and small molecules).

In conclusion, the well-differentiated propensities reported here for concordant and discordant states could thus be used to predict with high accuracy the most likely secondary structure of

any frequent HC, from the only information of a single sequence. This could be particularly useful for the characterization of orphan sequences, which can be highlighted through recently developed HCA-based tools [4,20]. HC reported here cover a large fraction of the HC present in globular domains (approximately 80 % of the total number of HC), frequently associated with α-helices and β-strands. Analysis of our SCOP95-derived database indicates that some HC, initially not present in the dictionary are now accessible to statistical analyses. These are longer than 7 amino acids, being associated with α-helices or corresponding to multiple clusters (associated with at least two SSR) (**Supplementary Data 9**). In the future, developing strategies for treating multiple HC as well as considering NMR data information, which was not primarily included in our reference datasets, should also allow increasing the number of informative HC types. NMR information could also be used for investigating further the link between local plasticity and amino acid composition in specific HC.

ACKNOWLEDGMENTS

This work was supported by the Institut du Cancer – France (INCA DI-REP and MYSM1) and the Agence Nationale de la Recherche – France (ANR CE10-0006-03). The authors thank Tristan Bitard-Feildel for critical reading of the manuscript.

Legends to Figures

Figure 1: *Difference between simple binary patterns and conditioned binary patterns (Hydrophobic Clusters – HC). 1 and 0 stand for hydrophobic and non-hydrophobic amino acids, respectively. The condition for a binary pattern to become a hydrophobic cluster (HC) is that at least four non-hydrophobic amino acids (or a single proline) separate two successive hydrophobic amino acids. This distance, called connectivity distance, is linked to the 2D support used for the sequence representation. The simple binary pattern can be included in larger binary patterns that, when corresponding to the HC limits, are associated with distinct secondary structures of length matching that of the HC* [12].

Figure 2: *Correspondence between the amino acid sequence, the binary code, the Q-code, the P-code, the 2D HCA plot and the observed 3D structure. The example shown here is of a part of human ketohexokinase (pdb 2hlz). The sequence is translated into a binary code (1 = strong hydrophobic amino acids (V, I, L, M, F, Y, W), 0 = any other amino acid, except for P (*)). The HCA binary code is derived from this simple binary code by considering a connectivity distance of 4 (four "0" or a proline separating two "1"). This HCA binary code can be decomposed into basic units (called Q (Quark)-codes – V, M, U, D) following the three axes found on the 2D plot (in green). The corresponding P (Peitsch)-codes are also indicated (see Material and Methods). In the 2D HCA plot, the sequence is written on a α-helical net that is unrolled and duplicated. The contours of hydrophobic amino acids are joined together to form clusters. Special symbols are used for proline (star), glycine (diamond), serine (dotted square) and threonine (square). Correspondences with SSR (shown above the sequence and on the 3D structure) are shown with colored labels.*

Figure 3: *Illustration of the periodicity of strong hydrophobic residues (represented as 1) in the regular secondary structures. The orange line materializes the protein-solvent interface. (a) Concordant α-helix and β-strand have all their hydrophobic positions oriented to the protein core, while (b) Discordant clusters expose some of their hydrophobic positions to the solvent.* *The polypeptidic path drawing was extracted from http://commons.wikimedia.org/wiki/File:Emergent_Homochirality.png.*

Figure 4: A. *Distributions of pairwise sequence identities between HC sharing identical P-codes (HC lengths 4+) in the SCOP95 and in the SCOP40 databases.* ***B.*** *Top: Distributions of pairwise sequence identities for HC with P-codes 153 (mainly associated with α-helices, left) and 85 (mainly associated with β-strands, right). Some bars are missing for some sequence identity ranges because they are not covered by the calculations (For HC with P-codes 153 (8 amino acids long), 3 or 4 identical positions give sequence identities of 37.5% and 50% respectively, therefore the range "40-50" is skipped). Bottom: some examples of HC sharing 0 % sequence identity with concordant behaviors (two first structures) and discordant behaviors (last structure relative to the two others).*

Figure 5: *WebLogo representations of the observed secondary structures and amino acid composition in two HC types typical of α-helices (P-153) and β-strands (P-85), in concordant and discordant states (P-153 (473 and 87 clusters) and P-85 (224 and 113 clusters)). The secondary structures are shown on the left side (DSSP assignment); the compositions in amino acids are represented in the middle (each of the 20 amino acids) and at right (20 amino acids grouped in categories according to their structural preferences, as defined from the data presented in Figure 6 (****A*** *(helix-forming) = A, L, M, E, Q, K, R;* ***E*** *(extended, strand-forming) = V, I, C, T;* ***W*** *(aromatic residues) = F,Y,W and* ***C*** *(coil-forming) = G, D, N, S; histidine (H) being neutral) and P is excluded from HC (see Material and Methods)).*

Figure 6: *Analysis of Q-codes associated with HC types. The total number of each of the four Q-codes (occurrences) are counted within the 180 HC types with SSR propensities for α-helix (H,h in dark and light blue respectively) and β-strand (E,e in dark and light green respectively). Upper and lower cases indicate strong and weak affinities for RSS, respectively. These are expressed as percentages of each HC type class (H,h,E,e) (top panel) and as percentages of Q-codes (V,M,U,D) (middle panel). The bottom panel represents the total number of Q-codes (occurrence) in HC types as a function of their lengths.*

Figure 7: *Propensities of the amino acids for the different secondary structures: 1) calculated over the whole protein sequences selected from SCOP95 (A) and SCOP40 (B) databases, 2) within the HC limits (note that proline is now excluded from calculations): global statistics (C), distinguishing the concordant/discordant states of HC with helix and strand preferences (D), 3) global statistics distinguishing two discordant states (Disc and Disc80) for HC with helix (E) and strand preferences (F). Discordant_80 adds an extra criterion of having more than 80% of the opposite RSS. Comparison of the distributions of the total number of each of the 19 amino acids in concordant and discordant states showed significant differences (see text).*

Supplementary Data

Supplementary Data 1 : *Occurrences and percentages of the 20 amino acids in the SCOP95- and SCOP40-derived datasets and the correlation between the two datasets.*

Supplementary Data 2: *HC databases, built from the SCOP95 and SCOP40-derived datasets. A) Global statistics of the HC occurrences. B) Mean HC occurrence per HC type for HC with propensities for α-helices and β-strands. C) Occurrence of HC types as a function of HC type length. D) Occurrence of HC as a function of HC length. Mean lengths (standard deviations) are indicated for the SCOP95 and SCOP40 databases.*

Supplementary Data 3: *Concordant and Discordant HC. A) Definition of concordant/discordant states relative to the behavior expected from the dictionary. The occurrences of concordant, discordant and intermediate HC are counted according to the following rules: for HC with helix and strand affinities, HC is described as Concordant if more than 80% or 60 % of the HC respects the predicted RSS in the dictionary; Discordant if the HC presents less than 20% of the predicted RSS in the dictionary and Intermediate for the other cases. B) Occurrences of concordant, discordant and intermediate HC in the SCOP95 and SCOP40 datasets. The two numbers designate the 80- and 60-levels, whereas values within brackets refers to the number of concerned HC types. Mean HC occurrences per HC type are shown in the bottom part of panel B. C) Mean of the ratios calculated per HC type between concordant and discordant clusters (mean Rcd). D) Distribution (frequencies) of the ratios calculated per HC type between concordant and discordant clusters (Rcd).*

Supplementary Data 4: *HCA plots of the 180 HC types analyzed in this study, classified by SSR affinities (h,H,e,E) and as a function of the increasing ratio of the number of concordant*

and discordant HC (Rcd) in the SCOP95 database (60 % level for definition of the concordance/discordance states). HC are represented by both their P-codes (top line) and Q-codes (bottom line). The colors indicate the range of Rcd values. Bar indicates the Rcd values of 1, before which there are more discordant than concordant HC per HC type. Note that the HC type with P-code 4371 (Q-code DDUV) has no discordant representative in our database. HC types corresponding to the Q-codes M, U and D are shaded in colors. The Q-code V is not present in this list as it is mainly associated with coils.

Supplementary Data 5: *Occurrences of HC with affinities for the helix (H,h) and strand (E,e) states, distinguishing between the concordant, discordant and intermediate states. Upper and lower cases indicate strong and weak affinities for RSS, respectively. These data were calculated for the SCOP95- and SCOP40-derived databases.*

Two levels are considered for the definition of the concordant state: 80-level (Conc-80) and 60-level (Conc-60) if more than 80% and 60% of the HC respects the RSS affinity predicted from the dictionary, respectively. The discordant state (Disc) is assigned if the HC presents less than 20% of the RSS affinity predicted from the dictionary, and the Intermediate state (Int-80 and Int-60) for the other cases.

Supplementary Data 6: *WebLogo representations, for the 180 HC types considered here, of the observed secondary structure and amino acid composition (each of the 20 amino acids considered and amino acids grouped according to their RSS affinities). Data can be found http://www.impmc.upmc.fr/~callebau/SD6_WL, including a README file for detailed explanations.*

Supplementary Data 7: *Logarithms of amino acid propensities for the three secondary structure states (Helix, Strand, Coil) calculated for the SCOP95- and SCOP40-derived databases (whole protein sequences),* ***B*** *(right) Couples of amino acid propensities for the concordant and discordant states for HC with affinities for* α*-helices (top) and* β*-strands (bottom).*

Supplementary Data 8*: Propensities of amino acids for the Concordant/Discordant states, distinguishing HC with strong (H, E) and weak (h,e) affinities for RSS. Top panel: propensities for HC with helix (left) and strand (right) propensities, regardless the secondary structure which is actually observed. Bottom panel: propensities analyzed according to the secondary structure which is actually observed.*

Supplementary Data 9 *Distribution, according to the HC length, of the number of HC types (colored in red; each HC types being defined by a unique binary pattern) that are not listed in the previously established dictionary* [23]*, due to a too low occurrence. The total number of HC (occurrences) for a specific length is reported in blue.*

References

1. Tanaka S, Isono K. Correlation between Observed Transcripts and Sequenced Orfs of Chromosome-Iii of Saccharomyces-Cerevisiae. Nucleic acids research 1993;21(5):1149-1153.
2. Fischer D, Eisenberg D. Finding families for genomic ORFans. Bioinformatics 1999;15(9):759-762.
3. Light S, Basile W, Elofsson A. Orphans and new gene origination, a structural and evolutionary perspective. Curr Opin Struc Biol 2014;26:73-83.
4. Faure G, Callebaut I. A comprehensive repertoire of foldable segments within genomes. PLoS Comput Biol 2013;in press.
5. Rost B. Twilight zone of protein sequence alignments. Protein Engineering 1999;12(2):85-94.
6. Fourty G, Callebaut I, Mornon JP. Characterization of non-trivial neighborhood fold constraints from protein sequences using generalized topohydrophobicity. Bioinform Biol Insights 2008;2:47-66.
7. Xiong H, Buckwalter BL, Shieh HM, Hecht MH. 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 1995;92(14):6349-6353.
8. Bellesia G, Jewett AI, Shea JE. Sequence periodicity and secondary structure propensity in model proteins. Protein Sci 2010;19(1):141-154.
9. Ventura S, Serrano L. Designing proteins from the inside out. Proteins 2004;56(1):1-10.

10. Gaboriaud C, Bissery V, Benchetrit T, Mornon JP. Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. FEBS Lett 1987;224:149-155.

11. Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon JP. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. Cell Mol Life Sci 1997;53(8):621-645.

12. Woodcock S, Mornon JP, Henrissat B. Detection of secondary structure elements in proteins by hydrophobic cluster analysis. Protein Eng 1992;5:629-635.

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

14. Callebaut I, Moshous D, Mornon JP, de Villartay JP. Metallo-beta-lactamase fold within nucleic acids processing enzymes: the beta-CASP family. Nucleic Acids Res 2002;30:3592-3601.

15. Callebaut I, Malivert L, Fischer A, Mornon JP, Revy P, de Villartay JP. Cernunnos interacts with the XRCC4 x DNA-ligase IV complex and is homologous to the yeast nonhomologous end-joining factor Nej1. J Biol Chem 2006;281:13857-13860.

16. Callebaut I, Mornon J. LOTUS, a new domain associated with small RNA pathways in the germline. Bioinformatics 2010;26:1140-1144.

17. Callebaut I, Mornon J. From BRCA1 to RAP1: a widespread BRCT module closely associated with DNA repair. FEBS Lett 1997;400:25-30.

18. Girault JA, Labesse G, Mornon JP, Callebaut I. The N-termini of FAK and JAKs contain divergent band 4.1 domains. Trends Biochem Sci 1999;24(2):54-57.

19. Bitard-Feildel T, Heberlein M, Bornberg-Bauer E, Callebaut I. Detection of orphan domains in Drosophila using "hydrophobic cluster analysis". Biochimie 2015.

20. Faure G, Callebaut I. Identification of hidden relationships from the coupling of Hydrophobic Cluster Analysis and Domain Architecture information. Bioinformatics 2013;in press.

21. Faure G, Revy P, Schertzer M, Londono-Vallejo A, Callebaut I. The C-terminal extension of human RTEL1, mutated in Hoyeraal-Hreidarsson syndrome, contains harmonin-N-like domains. Proteins 2014;82(6):897-903.

22. Rebehmed J, Revy P, Faure G, de Villartay JP, Callebaut I. Expanding the SRI domain family: a common scaffold for binding the phosphorylated C-terminal domain of RNA polymerase II. FEBS Lett 2014;588(23):4431-4437.

23. Eudes R, Le Tuan K, Delettre J, Mornon JP, Callebaut I. A generalized analysis of hydrophobic and loop clusters within globular protein sequences. BMC structural biology 2007;7:2.

24. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247(4):536-540.

25. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577-2637.

26. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res 2004;14(6):1188-1190.

27. Wathen B, Jia Z. Folding by numbers: primary sequence statistics and their use in studying protein folding. Int J Mol Sci 2009;10(4):1567-1589.

28. White SH, Jacobs RE. The evolution of proteins from random amino acid sequences. I. Evidence from the lengthwise distribution of amino acids in modern protein sequences. J Mol Evol 1993;36(1):79-95.

29. West MW, Hecht MH. Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. Protein Sci 1995;4(10):2032-2039.

30. Broome BM, Hecht MH. Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis. J Mol Biol 2000;296(4):961-968.

31. Schwartz R, Istrail S, King J. Frequencies of amino acid strings in globular protein sequences indicate suppression of blocks of consecutive hydrophobic residues. Protein Sci 2001;10(5):1023-1031.

32. Vazquez S, Thomas C, Lew RA, Humphreys RE. Favored and suppressed patterns of hydrophobic and nonhydrophobic amino acids in protein sequences. Proc Natl Acad Sci U S A 1993;90(19):9100-9104.

33. Wang W, Hecht MH. Rationally designed mutations convert de novo amyloid-like fibrils into monomeric beta-sheet proteins. Proc Natl Acad Sci U S A 2002;99(5):2760-2765.

34. Richardson JS, Richardson DC. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. Proc Natl Acad Sci U S A 2002;99(5):2754-2759.

35. Wathen B, Jia Z. Residue patterning in helix interiors. Biochem Cell Biol 2010;88(2):325-337.

36. Walther D, Argos P. Intrahelical side chain-side chain contacts: the consequences of restricted rotameric states and implications for helix engineering and design. Protein Eng 1996;9(6):471-478.

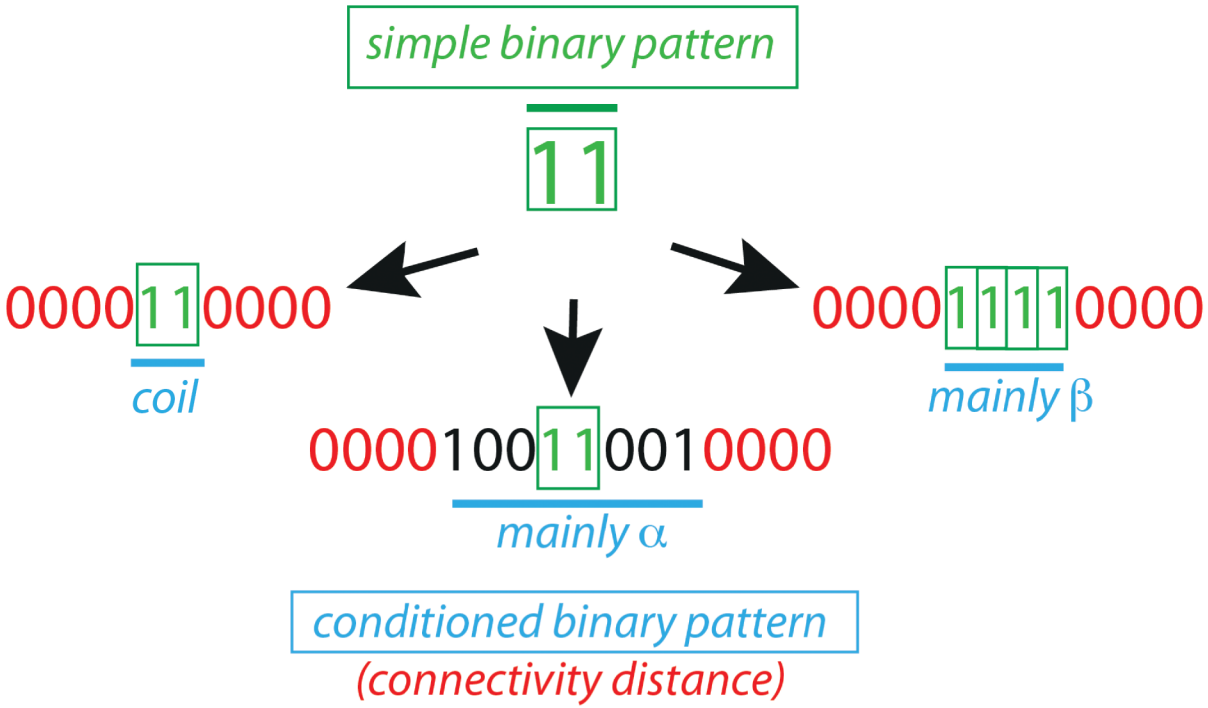

simple binary pattern
11
0000110000
coil
0000100110010000
mainly α
0000111110000
mainly β
conditioned binary pattern
(connectivity distance)

Figure 1

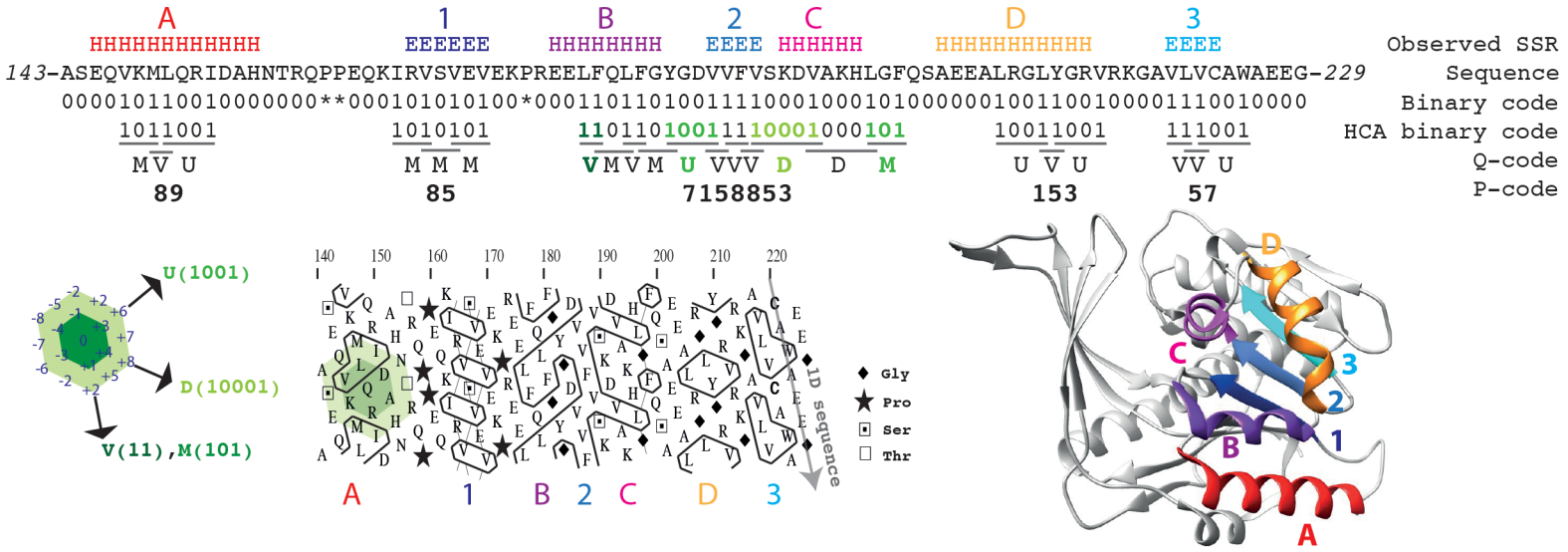

A 1 B 2 C D 3
HHHHHHHHHHHH EEEEEE HHHHHHHH EEEE HHHHHH HHHHHHHHHHH EEEE Observed SSR
143-ASEQVKMLQRIDAHNTRQPPEQKIRVSVEVEKPREELFQLFGYGDVVFVSKDVAKHLGFQSAEEALRGLYGRVRKGAVLVCAWAEEG-229 Sequence
0000101100100000000**000101010100*0001101101001111000100010100000010011001000011100100000 Binary code
1011001 1010101 11011010011110001000101 10011001 111001 HCA binary code
MV U M M M VMVM U VVV D D M U V U VV U Q-code
89 85 7158853 153 57 P-code
U(1001)
D(10001)
V(11),M(101)
140 150 160 170 180 190 200 210 220
1D sequence
Gly
Pro
Ser
Thr
A 1 B 2 C D 3
D
C
3
2
1
B
A

Figure 2

P-153 *(conditioned) binary pattern* P-85

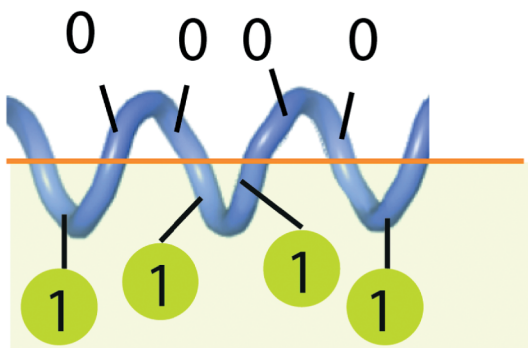
0000100110010000

Expected (pattern) : α helix

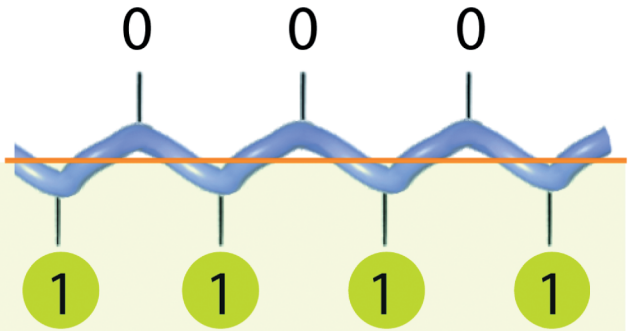
0000101010100000

Expected (pattern) : β strand

a) Concordance between the expected and observed 2D structures : Major state

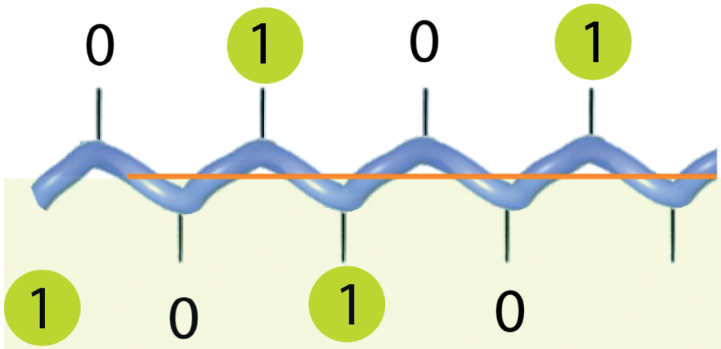


Observed : α helix

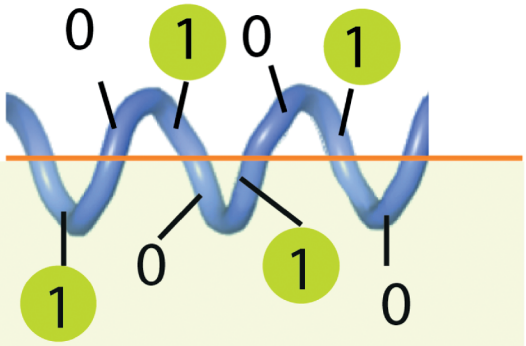


Observed : β strand

b) Discordance between the expected and observed 2D structures : Minor state



Observed : β strand



Observed : α helix

Figure 3

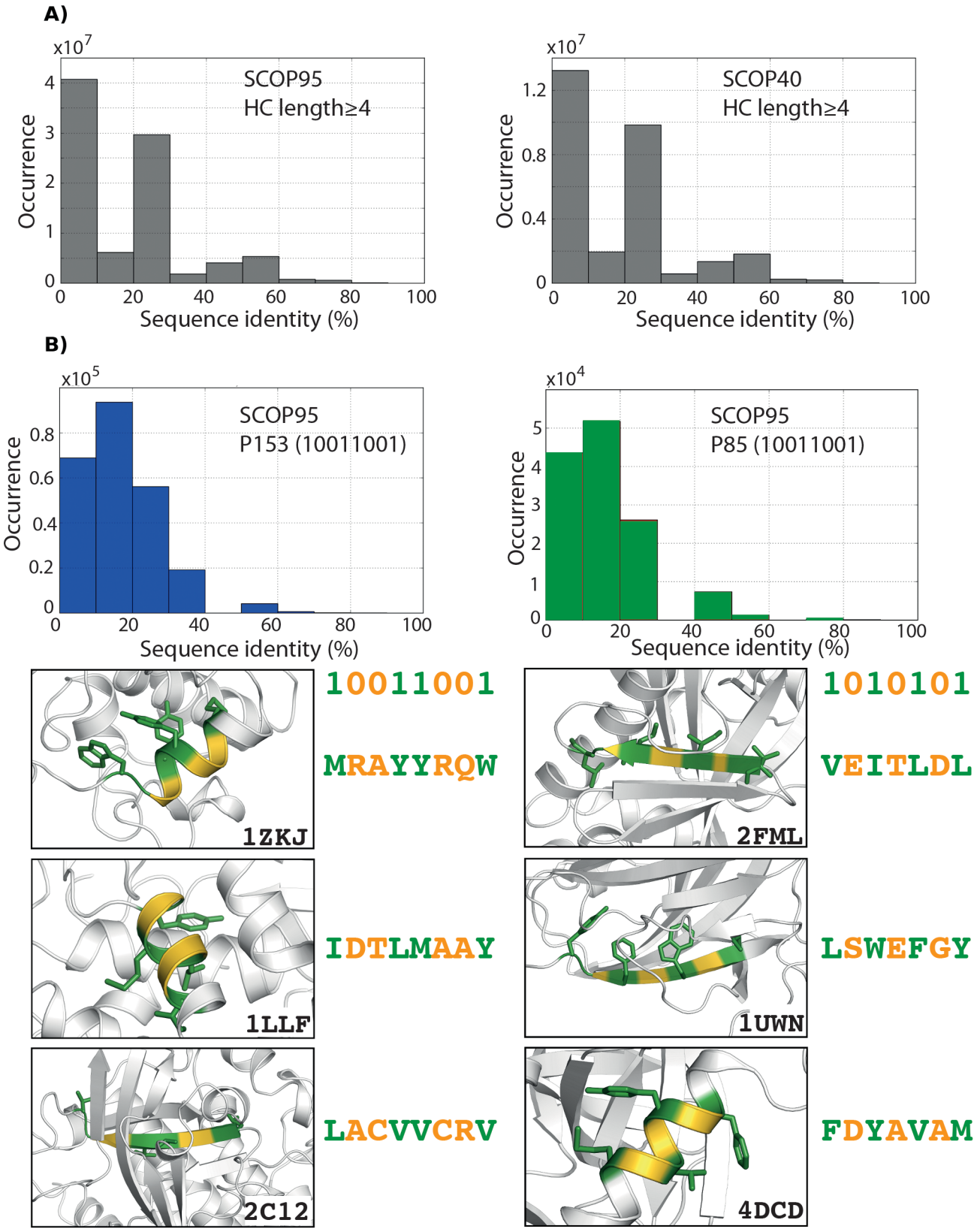

A)
x10^7
SCOP95
HC length≥4
Occurrence
Sequence identity (%)
x10^7
SCOP40
HC length≥4
Occurrence
Sequence identity (%)
B)
x10^5
SCOP95
P153 (10011001)
Occurrence
Sequence identity (%)
x10^4
SCOP95
P85 (10011001)
Occurrence
Sequence identity (%)
10011001
MRAYYRQW
1ZKJ
1010101
VEITLDL
2FML
IDTLMAAY
1LLF
LSWEFGY
1UWN
LACVVCRV
2C12
FDYAVAM
4DCD

Figure 4

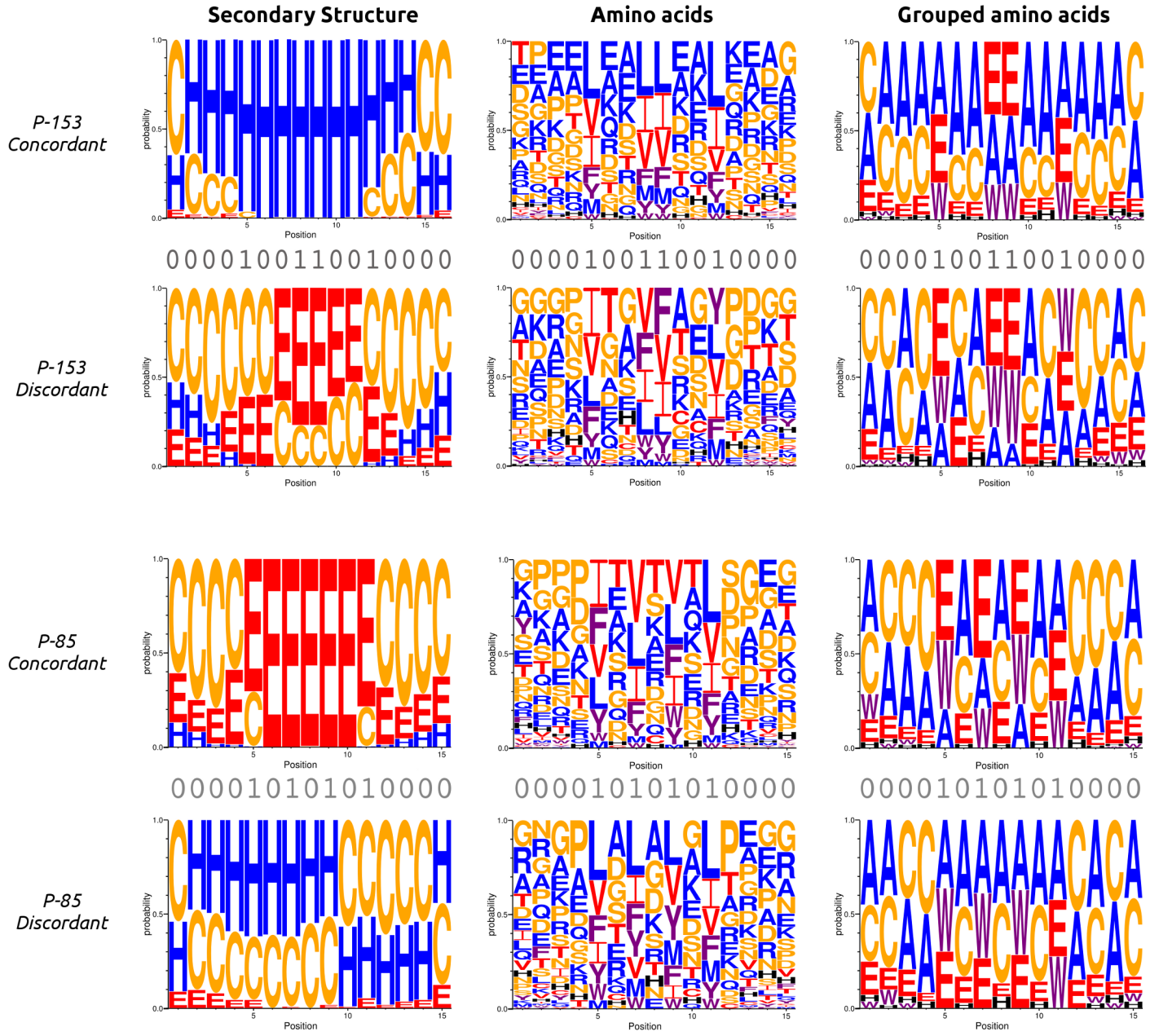

Secondary Structure
Amino acids
Grouped amino acids
P-153 Concordant
P-153 Discordant
P-85 Concordant
P-85 Discordant
0000100110010000
0000100110010000
0000100110010000
0000101010010000
0000101010010000
0000101010010000
probability
Position

Figure 5

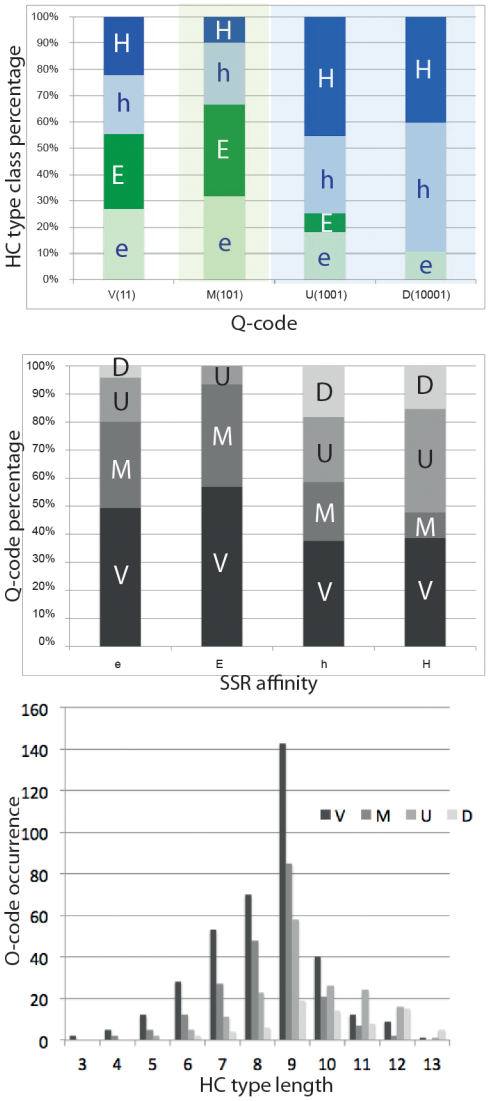

HC type class percentage
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
V(11)
M(101)
U(1001)
D(10001)
Q-code
Q-code percentage
e
E
h
H
SSR affinity
O-code occurrence
V
M
U
D
3
4
5
6
7
8
9
10
11
12
13
HC type length

Figure 6

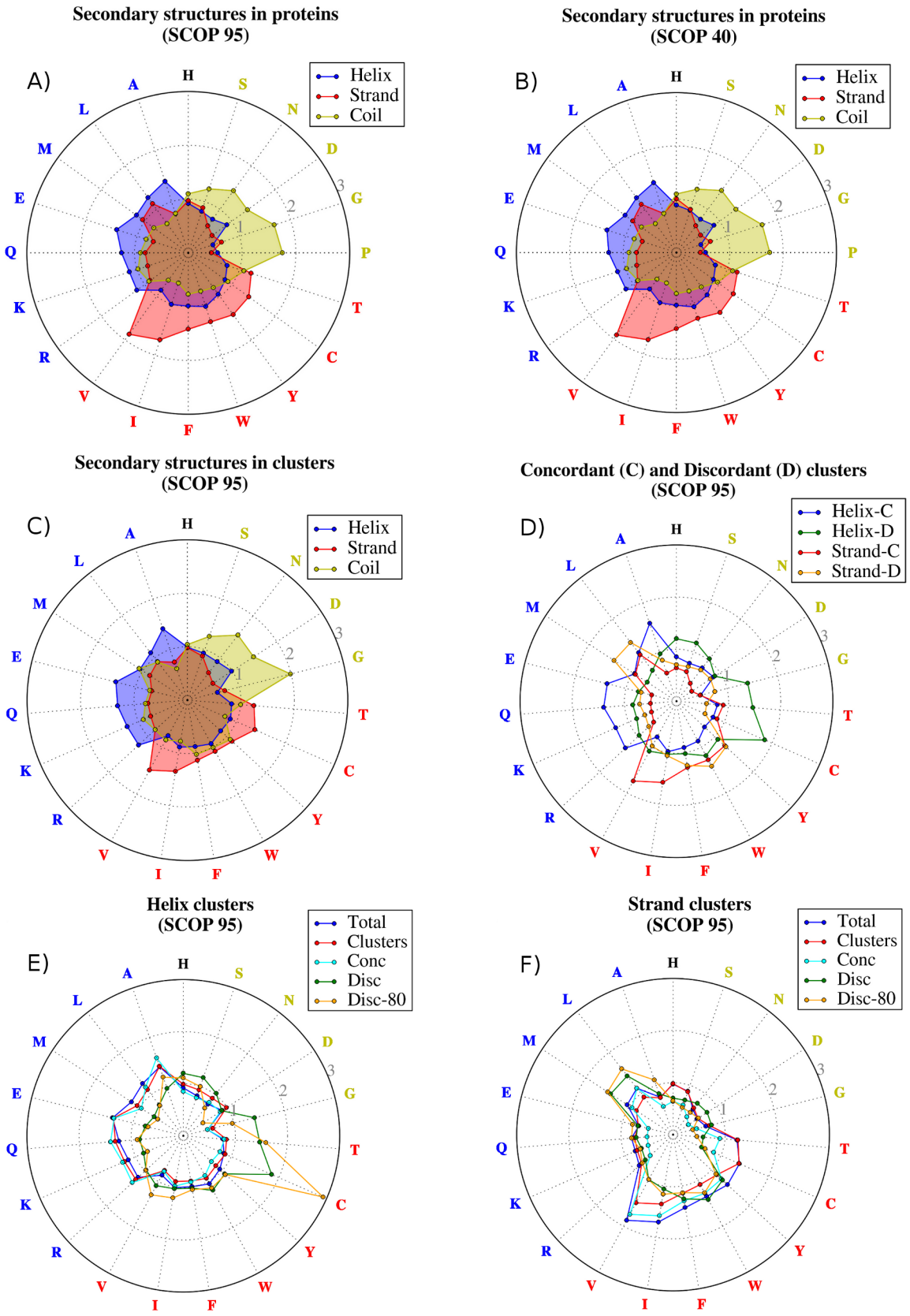

Secondary structures in proteins
(SCOP 95)
A)
Helix
Strand
Coil
Secondary structures in proteins
(SCOP 40)
B)
Helix
Strand
Coil
Secondary structures in clusters
(SCOP 95)
C)
Helix
Strand
Coil
Concordant (C) and Discordant (D) clusters
(SCOP 95)
D)
Helix-C
Helix-D
Strand-C
Strand-D
Helix clusters
(SCOP 95)
E)
Total
Clusters
Conc
Disc
Disc-80
Strand clusters
(SCOP 95)
F)
Total
Clusters
Conc
Disc
Disc-80
H S N D G P T C Y W F I V R K Q E M L A
1 2 3

Figure 7