



**HAL**  
open science

# Exploitation de signatures des repliements protéiques pour décrire le continuum ordre/désordre au sein des protéomes

Apolline Bruley

► **To cite this version:**

Apolline Bruley. Exploitation de signatures des repliements protéiques pour décrire le continuum ordre/désordre au sein des protéomes. Biochimie, Biologie Moléculaire. Sorbonne Université, 2022. Français. NNT : 2022SORUS474 . tel-04028741

**HAL Id: tel-04028741**

**<https://theses.hal.science/tel-04028741v1>**

Submitted on 14 Mar 2023

**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.

# SORBONNE UNIVERSITE

École doctorale 515 Complexité du Vivant  
*Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie*

---

## **Exploitation de signatures des repliements protéiques pour décrire le continuum ordre/désordre au sein des protéomes**

---

Par Apolline Bruley

Dirigée par Isabelle Callebaut et Elodie Duprat

Thèse présentée et soutenue publiquement le 05/12/2022

Devant un jury composé de :

Jean-Christophe Gelly	Maître de conférences, Université de Paris	Rapporteur
Raphaël Guérois	Directeur de recherche CEA, Université Paris Saclay	Rapporteur
Ingrid Lafontaine	Professeure, Sorbonne Université	Examinatrice
Anne Lopes	Maîtresse de conférences, Université Paris Saclay	Examinatrice
Lucie Bittner	Maîtresse de conférences, Sorbonne Université	Invitée
Isabelle Callebaut	Directrice de recherche CNRS, Sorbonne Université	Directrice de thèse
Elodie Duprat	Maîtresse de conférences, Sorbonne Université	Co-encadrante de thèse



Except where otherwise noted, this work is licensed under  
<http://creativecommons.org/licenses/by-nc-nd/3.0/>



# Remerciements

Je remercie en premier lieu l'école doctorale Complexité du Vivant d'avoir financé ma thèse, ainsi que l'ANR PHOSTORE qui m'a permis de la prolonger de quelques mois. Merci à Sorbonne Université et à l'IMPMC de m'avoir accueillie en leur sein.

Je tiens à remercier Jean-Christophe Gelly et Raphaël Guérois qui ont accepté de relire et d'évaluer ce manuscrit. Je remercie également Ingrid Lafontaine et Anne Lopes qui ont accepté d'examiner mon travail, ainsi que Lucie Bittner, qui a accepté d'être présente à ma soutenance et qui était là lorsque ce projet a été initié.

Je remercie également mes deux directrices de thèse Isabelle Callebaut et Elodie Duprat. Je vous remercie de m'avoir fait confiance durant le stage de M2 et durant ces trois années de thèse. Je vous remercie pour vos conseils et pour nos discussions scientifiques. Pendant ces presque quatre ans vous m'avez inspiré par votre passion, votre enthousiasme et votre rigueur, travailler avec vous a été un plaisir. Je vous remercie surtout de vous être toujours montrées compréhensives, disponibles et à l'écoute. Je vous remercie pour votre soutien et votre grande bienveillance.

Je tiens à remercier Tristan Bitard-Feildel qui a plus qu'initié une partie des travaux présentés ici. Je remercie également Maxime Millet, qui m'a aidée pour mes premiers pas sur un cluster de calcul et pour nos échanges sur les joies et (surtout) les frustrations de la bio-informatique.

Je remercie le personnel de l'IMPMC, et tout particulièrement celui du couloir 22-23, 5<sup>ème</sup> étage, pour son accueil et sa gentillesse.

Je remercie les doctorants, passés ou présents, de l'IMPMC que j'ai côtoyé pendant ma thèse, sans qui ces trois années auraient été bien différentes : Jeanne Caumartin, Dania Zuniga, Juliette Gaëtan, Geoffroy Gaschignard, Lucie Huart, Andréas Zoumpoulakis, Baptiste Truffet, Jaysen Sawmynaden, Léon Ambriambarijaona, Théo Magrino, Mohamad Harastani et les autres... Je tiens à remercier tout particulièrement mes premières co-bureau, Laura Galezowski et Cécile Bidaud. Merci pour votre bonne humeur et pour les longues discussions. Merci à toutes les deux d'avoir été là pour moi, même, et surtout, après votre départ. Merci à Cécile pour les séances de travail, chez elle ou en visio. Merci à Laura pour son écoute et ses conseils. Je remercie également Juliette Debrie : on aura fait nos thèses côte à côte du début à la fin, et nos longues sessions de travail en visio ont rendu cette période de rédaction moins solitaire et moins pénible, un énorme merci à toi. Je les remercie tous, pour les discussions, les déjeuners, les pauses café, les soirées et tant de moments inoubliables.

Je remercie Loriane, qui a toujours su se montrer présente et pleine d'encouragements malgré la distance qui nous sépare. Merci également à toi Maxime : cette thèse tu l'as un peu vécue avec moi, et tu sais mieux que personne ce qu'elle représente pour moi. Merci pour tes encouragements et merci de m'avoir soutenue jusqu'au bout. Merci surtout d'avoir toujours su me remotiver quand mon moral était au plus bas et d'avoir cru en moi, bien plus que moi-même.

Je remercie enfin ma famille. Je tiens, tout d'abord, à remercier chaleureusement Astrid pour sa présence, son soutien et sa générosité, de toujours mais de ces derniers mois tout particulièrement. Je remercie également mes parents, qui m'ont toujours encouragée et soutenue dans mes études, et qui m'ont toujours donné l'impression que rien n'était impossible. Sans vous je n'en serais pas là aujourd'hui, merci pour tout.

# Table des matières

<b>LISTE D'ABREVIATIONS .....</b>	<b>7</b>
<b>PREAMBULE.....</b>	<b>9</b>
<b>CHAPITRE 1. INTRODUCTION.....</b>	<b>11</b>
1.1. L'ORDRE DANS LES PROTEINES : RELATIONS SEQUENCE-STRUCTURE-FONCTION .....	13
1.1.1. Les quatre niveaux d'organisation de la structure des protéines .....	14
1.1.2. Évolution des séquences protéiques .....	16
1.1.3. Les domaines : unités structurales, évolutives et fonctionnelles des protéines .....	17
1.1.4. Séquences, structures, fonctions : ressources et outils .....	20
1.1.4.1. Les banques de données .....	20
1.1.4.2. Outils de recherche de similitude de séquences .....	21
1.1.4.3. Outils de prédiction des structures .....	22
1.1.4.4. Identification de signatures du repliement : Hydrophobic Cluster Analysis (HCA) .....	26
1.2. LES REGIONS INTRINSEQUEMENT DESORDONNEES .....	34
1.2.1. IDRs : une hétérogénéité de conformations.....	34
1.2.2. La flexibilité des IDRs : un avantage fonctionnel .....	35
1.2.2.1. Les fonctions des IDRs.....	35
1.2.2.2. Les sites fonctionnels d'interaction .....	36
1.2.3. Les séquences du désordre : biais et évolution .....	38
1.2.4. Les ressources et outils du désordre .....	39
1.2.4.1. Les banques de données du désordre .....	39
1.2.4.2. Les prédicteurs du désordre .....	41
1.2.4.3. AlphaFold2 et HCA, applicables à l'ordre et au désordre ?.....	42
1.3. LES INCONNUS DES PROTEOMES.....	45
1.3.1. Définitions et quantification .....	45
1.3.2. Causes de l'inconnu : erreurs d'annotation, désordre ou nouveauté ?.....	48
1.3.3. Ressources et outils.....	50
1.4. STRATEGIE ET OBJECTIFS DE LA THESE : .....	53
<b>CHAPITRE 2. DEVELOPPEMENT DU SCORE HCA ET COMBINAISON AVEC LES PREDICTIONS D'ALPHAFOLD2 POUR L'ETUDE DU CONTINUUM ORDRE/DESORDRE .....</b>	<b>56</b>
PRESENTATION DE L'ARTICLE.....	57
A SEQUENCE-BASED FOLDABILITY SCORE COMBINED WITH ALPHAFOLD2 PREDICTIONS TO DISENTANGLE THE PROTEIN ORDER/DISORDER CONTINUUM .....	61
Abstract .....	62
Introduction .....	63
Materials and Methods .....	65
Results.....	69
Discussion .....	76
References .....	82
Tables and Figures.....	88
Supplementary Information.....	97
<b>CHAPITRE 3. EXPLORATION DE L'INCONNU D'ALPHAFOLD2 : DU DESORDRE ET PLUS .....</b>	<b>122</b>
PRESENTATION DE L'ARTICLE.....	123
DIGGING INTO THE 3D STRUCTURE PREDICTIONS OF ALPHAFOLD2 WITH LOW CONFIDENCE: DISORDER AND BEYOND .....	127

<i>Abstract</i> .....	127
<i>1. Introduction</i> .....	127
<i>2. Material and Methods</i> .....	129
<i>3. Results</i> .....	131
<i>4. Discussion</i> .....	142
<i>References</i> .....	144
<b>CHAPITRE 4. DISCUSSION</b> .....	<b>150</b>
4.1. IDENTIFICATION DES SEQUENCES DE L'INCONNU .....	152
4.2. UTILISATION DU SCORE HCA POUR DECRIRE DE L'INCONNU .....	155
4.3. UTILISATION DU SCORE HCA POUR LA DESCRIPTION DE PROTEINES <i>DE NOVO</i> .....	161
<b>CHAPITRE 5. CONCLUSION ET PERSPECTIVES</b> .....	<b>168</b>
<b>BIBLIOGRAPHIE</b> .....	<b>173</b>



# Liste d'abréviations

**3D** : tridimensionnel

**aa** : acides aminés

**AF2** : AlphaFold2

**AFDB** : AlphaFold structure DataBase

**BLAST** : Basic Local Alignment Search Tool

**CDS** : Coding Sequence

**DUF** : Domains of Unknown Function

**HCA** : Hydrophobic Cluster Analysis

**HMM** : Hidden Markov Model

**IDD** : Intrinsically Disordered Domain

**IDP** : Intrinsically Disordered Protein

**IDR** : Intrinsically Disordered Region

**MoRFs** : Molecular Recognition Features

**PDB** : Protein Data Bank

**pLDDT** : predicted Local-Distance Difference Test

**PSSM** : Position Specific Scoring Matrix

**SCOPE** : Structural Classification of Proteins - extended

**SLiMs** : Short Linear Motifs

**SSR** : Structure Secondaire Régulière





# Préambule

Depuis le premier séquençage d'un génome en 1977 (bactériophage : 5 kb, 11 gènes) (Sanger et al., 1977), de nombreuses avancées ont été accomplies en termes de coût et d'efficacité, en particulier grâce à des développements expérimentaux (e.g., amplification de fragments d'ADN, séquençage de longues lectures via l'usage de nanopores, biologiques ou non) et de traitement numérique des lectures issues des séquenceurs (e.g., algorithmes de « binning » et d'assemblage de lectures). Ainsi, la méthode de séquençage de Sanger (Sanger et al., 1977), qui a permis la publication d'une première version de la séquence du génome humain en 2001 après 10 années d'efforts (Lander et al., 2001), a laissé place aux méthodes de nouvelles générations (Goodwin et al., 2016). Celles-ci permettent d'obtenir rapidement des séquences de génomes complets à partir de cellules uniques, mais également à partir d'échantillons environnementaux complexes, indépendamment de la mise en culture des micro-organismes, rendant ainsi accessible une part croissante de la diversité microbienne (Hug et al., 2016). De plus, ces méthodes permettent le séquençage de génomes de grande taille (e.g., *Paris japonica*, 150 Gb (Sun et al., 2022)). L'évolution rapide de la génomique rend<sup>1</sup> aujourd'hui accessible plus de 25 000 génomes d'eucaryotes, 40 000 génomes de procaryotes et 5 000 génomes de virus, ces effectifs étant en perpétuelle augmentation. De nombreux projets ont ainsi émergé pour séquencer les génomes de communautés issues d'échantillons océaniques (e.g., TARA ocean ; Sunagawa et al., 2015), d'écosystèmes terrestres (e.g., GEM ; Nayfach et al., 2021) ou encore des microbiomes de l'homme (HMP ; Lloyd-Price et al., 2017).

Les enjeux associés à cette abondance de données sont d'accéder à la diversité fonctionnelle et au potentiel métabolique des organismes, de décrire leurs interactions et leurs relations à leur environnement, ainsi que de décrire l'ensemble des processus d'évolution et d'adaptation à l'œuvre dans le vivant (Bernard et al., 2018; Thomas et Segata, 2019). Leurs implications sont nombreuses dans le domaine de la santé (en nous aidant, par exemple, à mieux comprendre l'origine de la pathogénicité de certaines bactéries (Melnyk et al., 2019), les processus d'antibiorésistance (Riesenfeld et al., 2004), ou encore la diversité et les rôles métaboliques des microbiomes humains (Lloyd-Price et al., 2017)) et de l'environnement (e.g. réponses des écosystèmes aux changements globaux (Reed et al., 2014)). L'exploitation de ces données permet ainsi l'identification d'innovations fonctionnelles au sein des génomes, qui présentent des intérêts pour la recherche fondamentale (e.g. découverte des systèmes CRISPR-Cas9 (Jinek et al., 2012)) et/ou la recherche appliquée (e.g. pharmacologie, technologies (biotransformations, biorémédiation)) (Bernard et al., 2018).

Le potentiel fonctionnel d'un organisme est conféré par l'expression de son protéome, c'est-à-dire l'ensemble des protéines pouvant être encodées par ses gènes. L'accès à la connaissance de ce potentiel repose sur les méthodes d'annotation automatique des génomes, consistant à identifier les régions codantes et à prédire leurs fonctions, respectivement sur la base des biais

---

<sup>1</sup> <https://www.ncbi.nlm.nih.gov/genome/browse/>.

dans l'usage du code génétique (Al-Turaiki et al., 2011) et des similitudes de leurs séquences avec celles dont la fonction est déjà connue par l'expérience (Jiang et al., 2016). De nombreuses limitations dans ces approches restent à dépasser, rendant actuellement inaccessible une part importante du potentiel fonctionnel des génomes séquencés et conduisant à l'émergence récente du concept de « matière noire » des génomes (Rinke et al., 2013).

Récemment, des techniques d'apprentissage profond (ou « deep learning », dénomination liée au nombre de couches dans les réseaux de neurones utilisés) ont permis d'utiliser l'information évolutive contenue dans ces abondantes données de séquences de protéomes afin de développer des prédicteurs particulièrement efficaces de la structure tridimensionnelle des protéines (Baek et al., 2021; Jumper et al., 2021). La base de données associée à l'outil AlphaFold2 rend désormais accessible les modèles de structures 3D pour l'ensemble des protéomes connus, soit actuellement plus de 200 millions de séquences de protéines. En accédant à l'information de structure des protéines, la promesse de repousser les limites de l'annotation fonctionnelle des protéines est faite. Cependant, une première étude réalisée sur les prédictions AlphaFold2 pour le protéome humain a montré qu'une partie seulement de la matière noire des protéomes pouvait être dévoilée. L'exploration de la part d'inconnu des protéomes reste plus que jamais d'actualité.

# **Chapitre 1. INTRODUCTION**

Les protéines, ou chaînes polypeptidiques, sont des polymères linéaires d'acides aminés, réunis par des liaisons covalentes appelées liaisons peptidiques. Dans nombre de cas, les protéines forment des ensembles hétérogènes en termes de structure, dans lesquelles alternent des régions ordonnées et des régions désordonnées (Figure 1.1). Ce chapitre est consacré à une brève introduction reprenant les concepts utilisés tout au long de mon travail et relatifs à ces deux types de régions et à leur annotation. En dernière partie de cette introduction, nous verrons qu'une part significative des protéomes reste à ce jour inconnue, que cette absence de connaissances concerne les relations évolutives, la structure et/ou la fonction des protéines. Je décrirai plus particulièrement les caractéristiques de ces séquences, qui constituent la matière noire des protéomes et les outils d'annotation récemment développés dans ce contexte.

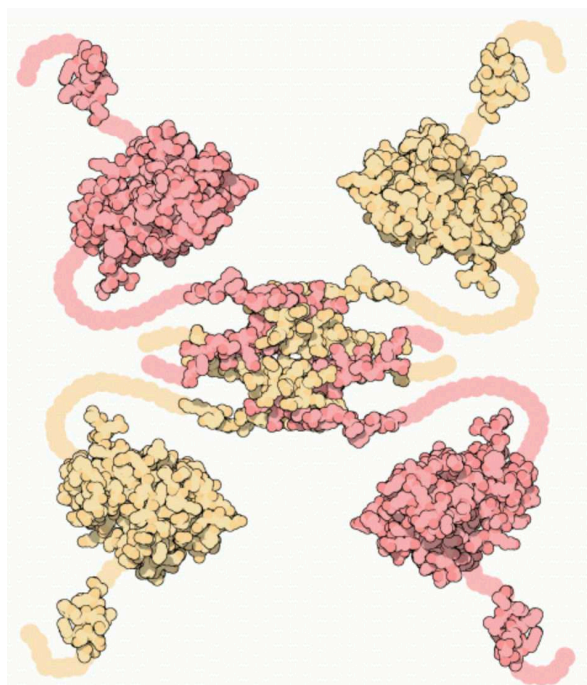


Figure 1.1. Illustration de l'hétérogénéité structurale de protéines multi-domaines, avec l'exemple de la protéine p53 (illustration extraite de pdb101 'molecule of the month'<sup>2</sup>), formée de 4 chaînes identiques (en jaune et en orange) composées chacune d'une alternance de portions flexibles et de portions structurées (domaines).

---

<sup>2</sup> <https://pdb101.rcsb.org/motm/31>

## 1.1. L'ordre dans les protéines : relations séquence-structure-fonction

Les cellules renferment de nombreux processus biologiques, portés par les protéines (Figure 1.2). Celles-ci sont de taille variable et sont retrouvées dans tous les compartiments cellulaires. Elles exercent des rôles aussi essentiels que divers, elles constituent par exemple la matrice qui maintient l'intégrité structurale des cellules, elles transmettent des messages au sein des cellules par des voies de signalisation, ou catalysent des transformations chimiques essentielles. C'est l'adoption d'une structure tridimensionnelle (3D) qui permet généralement aux protéines d'assurer des fonctions biologiques spécifiques, via les propriétés de sites fonctionnels spécifiques (e.g. enzymes), leurs propriétés de surface (e.g. forme, potentiel électrostatique) et/ou leur flexibilité.

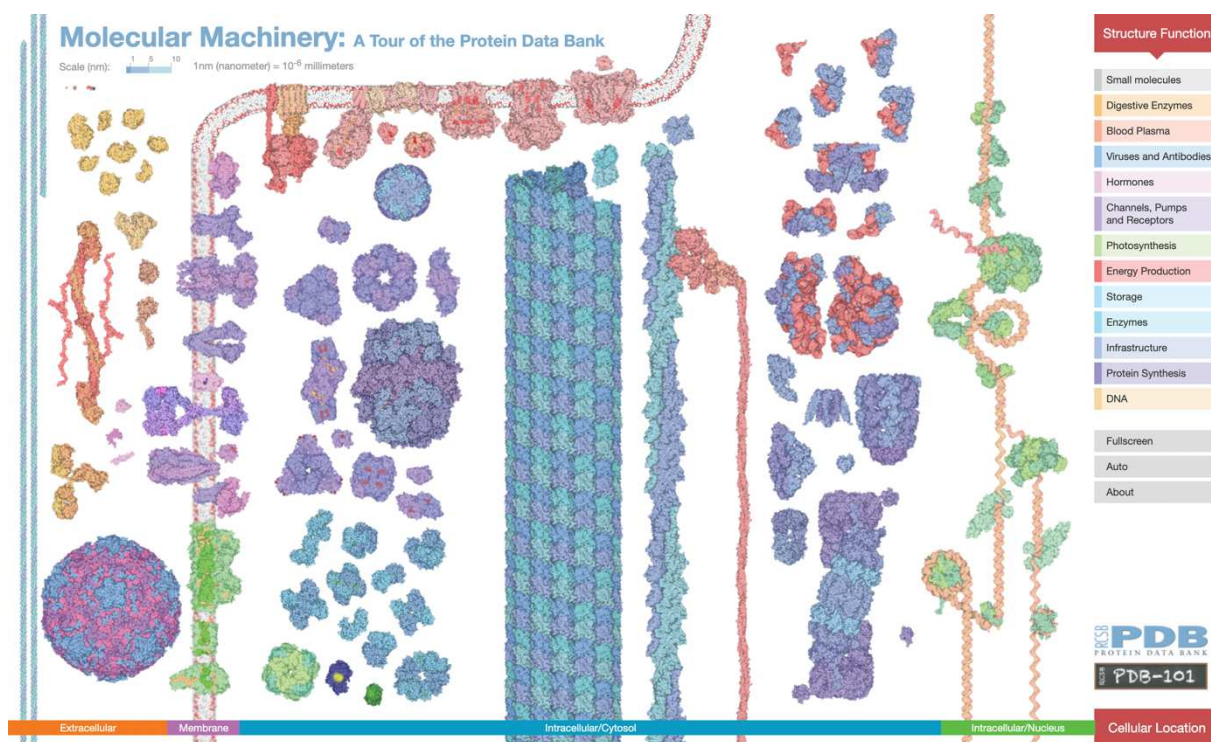


Figure 1.2. Illustration de la machinerie moléculaire et de ses structures protéiques. Illustration extraite de pdb101<sup>3</sup>.

<sup>3</sup> <https://cdn.rcsb.org/pdb101/molecular-machinery/>

### 1.1.1. Les quatre niveaux d'organisation de la structure des protéines

Il existe quatre niveaux de structuration des chaînes polypeptidiques. Le premier niveau, la **structure primaire**, n'est autre que la séquence en acides aminés, fruit de la traduction de l'ARN messager en protéines par le ribosome, depuis une extrémité N-terminale (terminaison amine) jusqu'à une extrémité C-terminale (terminaison carboxyle). Au nombre de vingt, les acides aminés se distinguent les uns des autres par la nature de leur chaîne latérale, qui leur confère des propriétés physico-chimiques spécifiques (Figure 1.3). Anfinsen et collaborateurs (Anfinsen et al., 1961) ont montré que l'intégralité de l'information nécessaire à l'adoption par une protéine de sa structure 3D est contenue dans sa séquence en acides aminés.

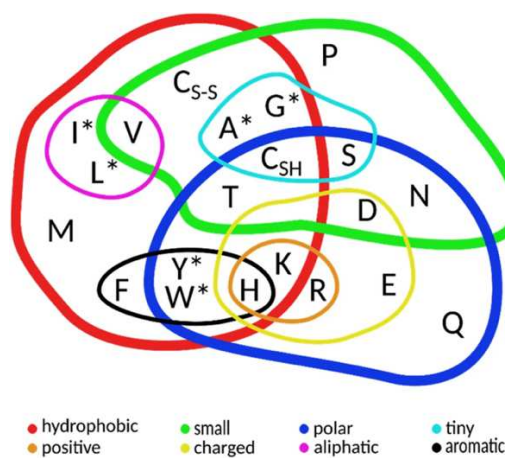


Figure 1.3. Classement des acides aminés en fonction de leurs caractéristiques physicochimiques. Les acides aminés sont représentés selon leur code à 1 lettre. C<sub>SH</sub> correspond à sa forme réduite de la cystéine et C<sub>S-S</sub> à la forme oxydée (ponts disulfures). Figure extraite de Then et al., 2020, qui suit la classification proposée par Taylor, 1986.

Le deuxième niveau, la **structuration secondaire**, décrit l'organisation locale de la chaîne polypeptidique. Les structures secondaires régulières (SSR) sont caractérisées par l'engagement des groupes amine (NH) et carboxyle (CO) de la chaîne principale dans des liaisons hydrogène. En neutralisant ainsi ces groupes polaires, le réseau de liaisons hydrogènes permet la formation de cœurs hydrophobes, propres au repliement des domaines globulaires de protéines (voir ci-après). Les deux types principaux de structures secondaires régulières, dont l'existence a été prédite par Pauling et Corey (Pauling et Corey, 1951), sont les hélices alpha et les brins beta (Figure 1.4). Il existe plusieurs programmes permettant l'attribution des structures secondaires à partir des coordonnées atomiques des structures 3D (Kabsch et Sander, 1983; Labesse et al., 1997; Richards et Kundrot, 1988; Sklenar et al., 1989), selon différents critères. Le plus communément utilisé, DSSP (Kabsch et Sander, 1983), propose, à partir de l'analyse du réseau de liaisons hydrogène, une classification en 8 états, distinguant trois types d'hélices, deux types de brins beta et trois types de boucles (éléments liant les structures secondaires entre elles). Partant de l'observation que les structures secondaires sont plus conservées que ne le sont les séquences (Mizuguchi et Go, 1995) ce programme d'attribution,

ainsi que des programmes de prédiction de structures secondaires (e.g. Buchan & Jones, 2019; Drozdetskiy et al., 2015), sont largement utilisés afin d'améliorer les correspondances dans les processus d'alignement de séquences (voir ci-après).

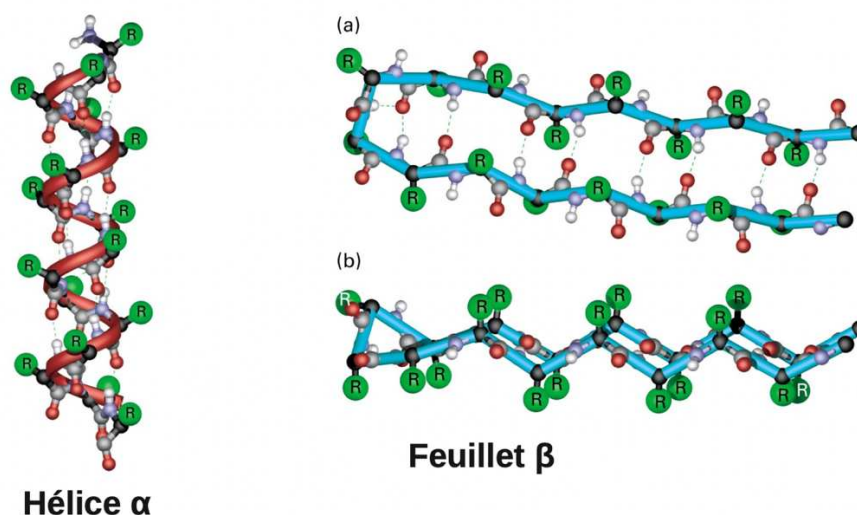


Figure 1.4. Réseaux d'acides aminés spécifiques aux hélices alpha (à gauche) ou aux feuillets beta (à droite)<sup>4</sup>.

Le troisième niveau, la **structure tertiaire**, se rapporte à l'agencement des éléments de structure secondaire en une structure 3D compacte. Le repliement 3D de la chaîne polypeptidique en cette structure compacte résulte des interactions formées par les acides aminés entre eux (pouvant impliquer les atomes des chaînes principales comme latérales) et avec le milieu. Selon leurs propriétés physico-chimiques (Figure 1.3), les acides aminés forment différents types d'interactions. La force motrice du repliement des protéines globulaires solubles est générée par les chaînes latérales hydrophobes des acides aminés qui fuient l'eau en s'enfouissant au cœur de la protéine, formant ainsi un cœur hydrophobe (Pace et al., 2014) et réduisant l'entropie conformationnelle. Les protéines membranaires se replient dans la membrane phospholipidique des cellules, les chaînes latérales des résidus hydrophobes, plus nombreux que dans les protéines solubles, sont donc très largement orientées vers les lipides (Hong, 2014). Les liaisons hydrogènes formées par les acides aminés polaires jouent également un rôle stabilisateur très important (Pace et al., 2014). Les interactions de van der Waals, les liaisons ioniques entre résidus chargés et les liaisons covalentes (comme les ponts disulfures formés par les cystéines), contribuent également au repliement. Les structures 3D sont déterminées expérimentalement par cristallographie aux rayons X, par spectroscopie de résonance magnétique nucléaire (RMN) ou par cryo-microscopie électronique (cryo-EM), les

<sup>4</sup><https://wiki.ubuntu.com/kmezoud/Bioinformatics?action=AttachFile&do=get&target=structure-3D-prediction.pdf>



coordonnées des atomes de ces structures sont recensées dans la Protein Data Bank (PDB (Berman et al., 2003)).

Des niveaux intermédiaires entre structures secondaires et tertiaires ont été décrits par certains auteurs, en particulier en utilisant des alphabets structuraux, conduisant à la constitution de bibliothèques de fragments structuraux permettant de décrire la diversité conformationnelle présente au sein des structures de protéines (Joseph et al., 2010; Kolodny et al., 2002). Il existe plusieurs alphabets structuraux qui se distinguent par le nombre de lettres structurales, leur longueur (en acides aminés) et la méthode utilisée pour les générer (Benros et al., 2006; Camproux et al., 1999; de Brevern et al., 2000; Dong et al., 2008; Kolodny et al., 2002; Ku et Hu, 2008).

Enfin, le quatrième niveau est celui de la **structure quaternaire**, se rapportant aux assemblages que les chaînes polypeptidiques peuvent former entre elles. Ces assemblages sont qualifiés d'homo- ou hétéro-oligomères, mais peuvent être aussi décrits vis-à-vis de leurs cinétiques d'interaction (complexes transitoires ou permanents) ou par leur caractère obligatoire ou non obligatoire (Liljas et al., 2017).

### 1.1.2. Évolution des séquences protéiques

L'évolution des séquences protéiques résulte d'évènements qui interviennent à l'échelle des séquences nucléotidiques : mutations ponctuelles (substitutions, délétions et insertions) ou réarrangements de plus grande taille (schématisés Figure 1.5) tels que l'émergence *de novo* ou la perte de gènes (Albalat et Cañestro, 2016) ou de régions, leur duplication, leur fusion ou leur fission. Ces évènements peuvent modifier les propriétés physico-chimiques des protéines codées par les gènes concernés, et affecter leurs structure 3D (interactions intra-moléculaires), leurs propriétés de surface et/ou leurs fonctions (Fowler et Fields, 2014; Jasinska et al., 2020). Cependant, des mutations compensatoires localisées dans le voisinage 3D des résidus concernés peuvent permettre le maintien de ces propriétés (Figure 1.6) (Altschuh et al., 1987; Levin et Mishmar, 2017; Marks et al., 2011). La co-évolution entre résidus d'une protéine peut donc refléter un voisinage 3D.

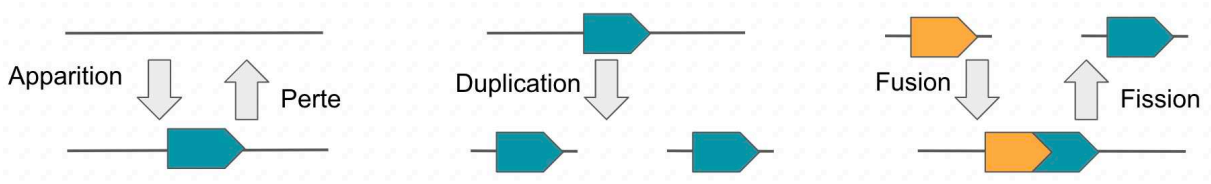


Figure 1.5. Processus d'évolution des répertoires génomiques. Figure adaptée de Wu et al., 2012.

Les séquences qui ont évolué à partir d'un ancêtre commun sont appelées séquences homologues. Il existe différents types d'homologues, détaillés ci-après. Les orthologues sont issus d'un événement de spéciation. Ils assurent une fonction similaire dans des organismes différents et présentent un niveau de conservation élevé (Koonin, 2005). Les paralogues sont issus d'un événement de duplication au sein d'un organisme donné, et constituent donc une forme de redondance fonctionnelle à l'issue de cet événement. Alors qu'une des copies pourra continuer à exercer la fonction de leur ancêtre, l'autre pourra diverger et acquérir une nouvelle fonction (néo-fonctionnalisation) ou devenir non fonctionnelle (pseudogénéisation) (Bershtein et al., 2021; Gabaldón et Koonin, 2013; Soria et al., 2014). En parallèle de ces mécanismes d'évolution verticale, des transferts de matériel génétique entre organismes peuvent se produire. Ces événements de transferts horizontaux sont courants chez les procaryotes, via des intermédiaires viraux (Weinbauer et Rassoulzadegan, 2004). Lorsque ces transferts impliquent des régions codantes, ils conduisent à l'observation d'un autre type de gènes homologues nommés xénologues. Ces séquences évoluent de façon autonome dans chacun de leurs hôtes.

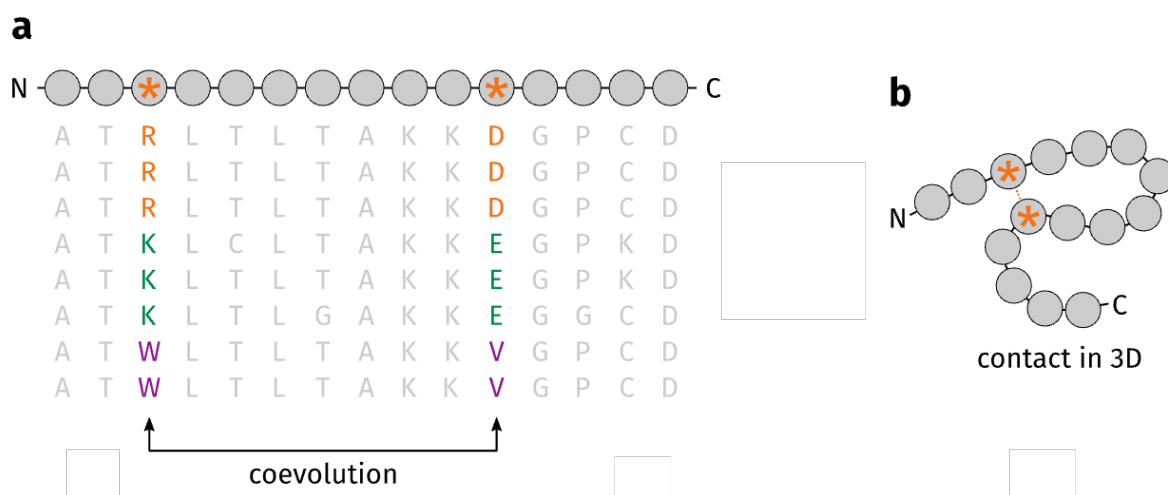


Figure 1.6. Illustration du phénomène de co-évolution entre deux acides aminés. a) Représentation d'un alignement multiple de séquences homologues, où deux positions qui ont co-évolué sont mises en avant. b) Représentation schématique de la structure d'une des protéines reprises dans l'alignement de séquences, dans laquelle est mise en évidence l'interaction entre les résidus qui ont co-évolué, permettant le maintien de la structure (pont salin ou interaction de van der Waals). Figure adaptée de Bittrich et al., 2019.

### 1.1.3. Les domaines : unités structurales, évolutives et fonctionnelles des protéines

La notion de domaines, introduite en 1973 (Wetlaufer, 1973), répond à une des trois définitions, qui généralement se recoupent : (i) une région qui se structure indépendamment du reste de la protéine, (ii) un segment qui évolue indépendamment du reste de la protéine, (iii)

une unité capable d'avoir une fonction indépendamment du reste de la protéine (Branden et Tooze, 1996) (Figure 1.7). La taille d'un domaine est en moyenne de 100 acides aminés, et peut varier entre environ 30 et 500 acides aminés (Wheelan et al., 2000).

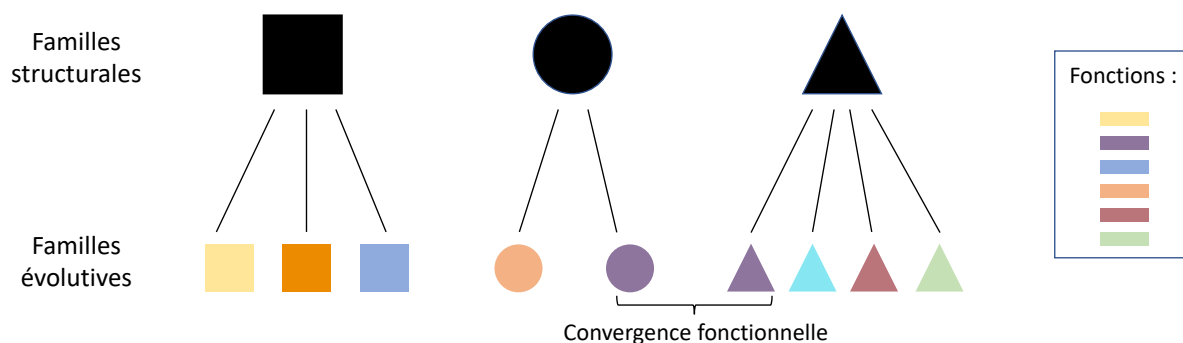


Figure 1.7. Lien entre familles structurales, familles évolutives et fonctions.

Cyrus Chothia en 1992 (Chothia, 1992), puis d'autres auteurs (Bordin et al., 2021; Holm et Sander, 1996; Kolodny et al., 2013; Koonin et al., 2002; Wolf et al., 2000; Zhang et DeLisi, 1998) ont montré qu'il n'existe qu'un nombre limité de repliements 3D que peuvent adopter les domaines. Ce nombre a été estimé à quelques milliers. Les domaines sont découpés et classés en familles structurales (repliements) dans des banques dédiées sur la base de la composition en structures secondaires et l'agencement de celles-ci (SCOPe (Andreeva et al., 2020), CATH (Redfern et al., 2007; Sillitoe et al., 2021), ECOD (Cheng et al., 2014)) (Figure 1.8, flèche bleue). La dernière version de la banque de classification structurale SCOPe (Tableau 1.1) (v2.08, sept 2021 (Chandonia et al., 2022)) que nous avons utilisée dans ce travail de thèse, recense 1256 repliements distincts. Au sein d'une même famille structurale, ces approches définissent différentes familles de séquences sur la base de leurs relations évolutives (Figure 1.8, flèche rouge). Dans le cas de SCOPe, 5084 familles évolutives sont ainsi définies (Tableau 1.1). Il est important de remarquer que ces banques contiennent relativement peu de protéines membranaires par rapport à l'ensemble des structures déterminées expérimentalement et répertoriées dans la PDB. Une annotation plus précise de ces protéines membranaires est proposée par la banque OPM (Lomize et al., 2006). Les domaines membranaires ont la particularité d'être composés uniquement d'un type de structures secondaires : une ou plusieurs hélices alpha (Figure 1.9.A) ou un ensemble de brins beta organisés en tonneaux (*beta-barrels*, Figure 1.9.B).

Par ailleurs, différentes bases de données de séquences de domaines ont été construites indépendamment de l'information structurale. Chaque famille de séquences homologues, considérées comme portant une fonction similaire, est représentée sous forme d'un profil probabiliste (PSSM ou HMM). Les deux banques les plus larges sont CDD et InterPro (Scaiewicz et Levitt, 2015), et proposent toutes deux des outils permettant l'identification de nouveaux membres (Jones et al., 2014; Marchler-Bauer et Bryant, 2004). Cette approche est à la base de l'annotation fonctionnelle que j'exploite dans mes travaux de thèse. InterPro (InterPro 90.0, 4 août 2022 (Blum et al., 2021)) recense 11954 familles de domaines.

## 1.1. L'ordre dans les protéines : relations séquence-structure-fonction

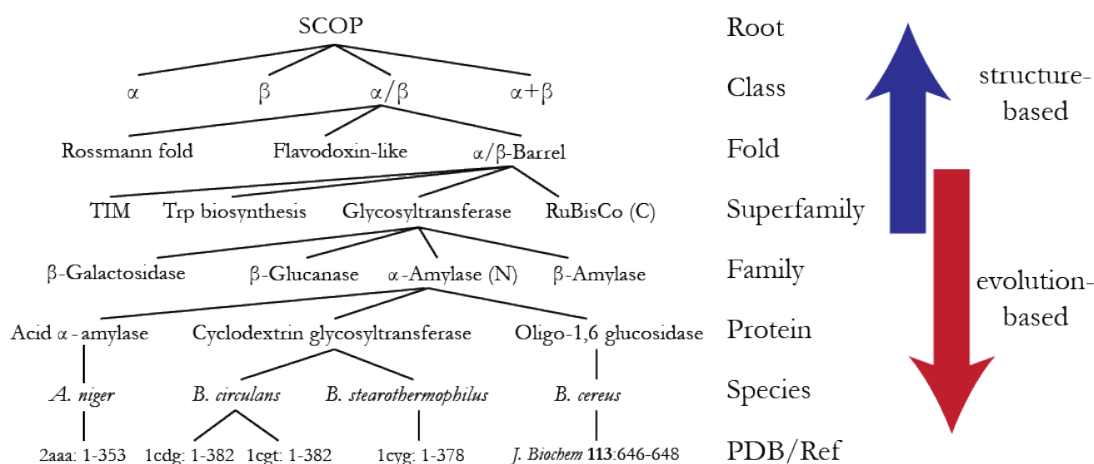


Figure 1.8. Schéma de classification de la banque de donnée SCOPe.

Classes	Nombre de folds	Nombre de superfamilles	Nombre de familles
<b>a : Protéines all- alpha</b>	290	519	1089
<b>b : Protéines all- beta</b>	180	375	993
<b>c : Protéines alpha et beta (a/b)</b>	148	247	1003
<b>d : Protéines alpha et beta (a+b)</b>	396	580	1387
<b>e : Protéines multi-domaines (alpha et beta)</b>	74	74	128
<b>f : Protéines et peptides de la membrane et de la surface cellulaire</b>	69	131	204
<b>g : Petites protéines</b>	100	141	280
<b>Totaux</b>	<b>1257</b> (26 nouvelles)	<b>2067</b> (42 nouvelles)	<b>5084</b> (88 nouvelles)

Tableau 1.1. Statistiques au 28 juillet 2021 de la SCOPe 2.08, recensant 106976 entrées PDB et 344851 domaines. SCOPe, consultée le 9 octobre 2022.

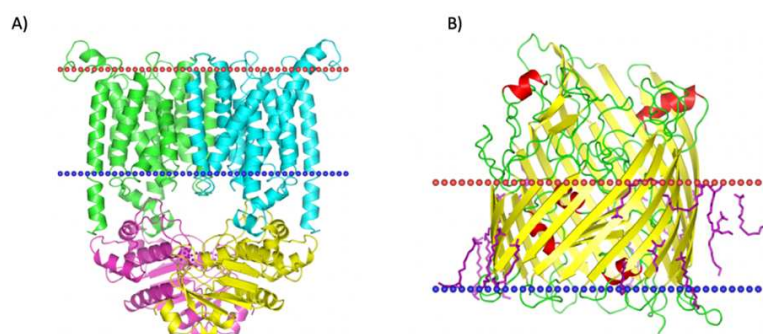


Figure 1.9. Exemples de domaines membranaires. La localisation prédite de la membrane est schématisée par les lignes rouges et bleues. A) Domaine composé de plusieurs hélices transmembranaires (transporteur BtuCD d'*E. coli* (superfamille ABC, transport de la vitamine B12), pdb : 1L7V). B) Domaine composé de brins beta formant un tonneau beta (transporteur de la membrane externe d'*E. coli* FecA, pdb : 1KMO). Illustrations extraites de la banque OPM.

La hiérarchie présente dans les banques de classification structurales permet donc de distinguer les notions de familles de repliement (agencement et connexions similaires de structures secondaires régulières) et familles de domaines (ensemble de séquences de même repliement présentant des relations évolutives et une fonction commune). Ainsi, une famille de repliement peut regrouper différentes familles de domaines (on parle alors de convergence structurale, c'est-à-dire sans lien évolutif), dont certaines sont particulièrement aptes à être retrouvées dans différentes architectures (Basu et al., 2009) (Figure 1.7). La combinatoire de l'ensemble des domaines rend ainsi compte de la diversité des séquences à l'échelle des protéomes (Scaiewicz et Levitt, 2018).

### 1.1.4. Séquences, structures, fonctions : ressources et outils

#### 1.1.4.1. Les banques de données

Les séquences, domaines et structures des protéines, ainsi que les fonctions qui leur ont été assignées (ou l'absence de fonction) sont consignés dans de nombreuses bases de données, jouant un rôle important dans la centralisation de ressources exhaustives permettant l'annotation des protéines. Les données sont des séquences protéiques entières (e.g. UniProt (The UniProt Consortium, 2021), banque non redondante (nr) du NCBI (Sayers et al., 2021), Big Fantastic Database (BFD) (Steinegger et al., 2019b; Steinegger et Söding, 2018)), des profils de séquences entières (e.g. eggNOG regroupant les séquences orthologues dans des profils uniques (Huerta-Cepas et al., 2019)), des profils de séquences de domaines (e.g. InterPro (Blum et al., 2021); Conserved Domain Database (CDD) (Lu et al., 2017), Pfam (Mistry et al., 2021)) ou enfin les structures 3D déterminées expérimentalement (e.g. PDB (Berman et al., 2003)), qui peuvent être découpées en domaines structuraux (e.g. SCOPe (Chandonia et al., 2022) et CATH (Sillitoe et al., 2021)) ainsi que les structures 3D prédites (e.g. AFDB, prédictions par AlphaFold2 (Varadi et al., 2022)). Certaines de ces ressources proposent une classification des données : CATH et SCOPe sur la base des structures (comme vu ci-dessus) ; ECOD, InterPro, CDD et SUPERFAMILY sur la base de l'information évolutive.

InterPro, par exemple, propose de classer les domaines en superfamilles homologues (S), familles (F), domaines (D), et les domaines eux-mêmes peuvent être catalogués selon leur fonction et caractéristiques comme des répétitions (R), ou des sites (S) d'intérêts (sites actifs, sites de liaison, site conservés, sites transmembranaires).

La banque de données OPM, que j'ai également exploité dans le cadre de cette thèse, regroupe les protéines membranaires présentes dans PDB, et les classe selon six critères différents dont leur composition en structures secondaires (e.g. domaine membranaire composé d'hélices alpha polytopique (plusieurs passages), bitopique (un seul passage), ou domaine composé de brins beta organisés en tonneau beta) ou leur localisation (e.g. membrane bactérienne interne gram-négative, membrane plasmique d'eucaryote, membrane de lysosome)

(Lomize et al., 2006). Shimizu et al. (2018) ont montré qu'OPM est la banque de données de structures 3D de domaines membranaires détenant le plus grand nombre d'entrées (9842, contre 4229 pour mpstruc (White, 2009), 3104 pour PDBTM (Tusnady et al., 2004) et 1457 pour SCOPE, au moment de l'étude).

Dans ces banques de données, on retrouve une annotation fonctionnelle et/ou structurale associée aux protéines ou aux domaines. Certaines entrées restent cependant de fonction inconnue, elles porteront l'annotation *unknown function* ou DUF (*Domain of Unknown Function*) (Bateman et al., 2010). D'autre part, l'étiquette *hypothetical proteins* regroupe les protéines prédites à partir de données de séquençage, protéines putatives et protéines non caractérisées (Sahoo et al., 2020).

#### 1.1.4.2. Outils de recherche de similitude de séquences

Pour annoter une séquence, la méthode traditionnelle consiste à rechercher des parentés dans les banques de données de séquences et/ou de domaines, et à lui transférer leur(s) annotation(s). Ces parentés ou relations d'homologie sont prédites sur la base de la recherche de similitudes.

De nombreux programmes existent pour la recherche de similitude dans une banque de séquences, et peuvent être appliqués à de grands jeux de données grâce au développement d'algorithmes rapides et efficaces, tels que diamond (Buchfink et al., 2015), qui correspond à une version accélérée de l'algorithme BLAST (Camacho et al., 2009). Cet algorithme est une heuristique, dont le fonctionnement est basé sur le découpage des séquences en mots de taille  $k$ , ainsi les séquences à annoter ne seront alignées qu'aux séquences de la banque de données partageant des mots similaires.

D'autres programmes se concentrent sur la détection d'orthologues. Ces méthodes permettent de regrouper des séquences similaires, permettant ainsi l'annotation rapide de celles-ci. Parmi ces algorithmes de classification non supervisée, on retrouve orthoMCL (Fischer et al., 2011), INPARANOID (Remm et al., 2001) ou MMseqs2 (Steinegger et Söding, 2018). Ce dernier algorithme est particulièrement adapté aux très grands jeux de données de séquences, une comparaison des séquences par paires en se basant sur leur composition en mots de taille  $k$  lui permet de générer des graphes de séquences dans lesquels les séquences similaires sont reliées et à partir desquels il définit des groupes de séquences similaires.

Considérer l'information évolutive, reprise sous forme de profils ou de modèles de Markov cachés (HMM), permet d'augmenter la sensibilité de détection d'homologues. Les algorithmes de recherche itérative (PSI-BLAST (Altschul et al., 1997) et HHblits (Steinegger et al., 2019a)) permettent la construction d'un profil à partir de la séquence à annoter, complétée par des recherches de similitudes itératives avec les nouveaux homologues identifiés. A chaque itération, le profil sera donc de plus en plus complet, et la recherche permettra de détecter, par recherche dans des banques de séquences, des relations d'homologie de plus en plus éloignées.

D'autres algorithmes permettent une recherche à partir d'un profil dans des banques de profils (e.g. HHsearch (Steinegger et al., 2019a) et hmmscan (Eddy, 2011)), pour la mise en évidence de relations d'homologie alors même que les séquences partagent moins de 20% d'identité (Blake et Cohen, 2001; Steinegger et al., 2019a).

Les méthodes décrites ci-dessus reposent toutes sur des alignements séquences-séquences ou séquences-profils, une étape dont s'affranchissent les méthodes basées sur les modèles de langage des protéines (ou pLM pour « protein Language Model ») (Bepler et Berger, 2019; Elnaggar et al., 2022; Rives et al., 2021; Schütze et al., 2022). Ces méthodes apparues récemment utilisent des algorithmes d'apprentissage tirés du traitement automatique du langage naturel. Dans le cadre des protéines, chaque acide aminé est considéré comme un mot, et est encodé en fonction du contexte dans lequel il est retrouvé (on parle d'« embedding »). Ces méthodes ont des résultats prometteurs, elles permettent de détecter des relations d'homologie entre séquences qui partagent moins de 20% de pourcentage d'identité (Heinzinger et al., 2022; Littmann et al., 2021).

#### *1.1.4.3. Outils de prédiction des structures*

Jusqu'à récemment, les méthodes de prédictions de structure 3D les plus efficaces en terme de précision se limitaient à une modélisation basée sur l'homologie, plus justement dénommée modélisation comparative (Jisna et Jayaraj, 2021). Des outils comme Phyre2 (Kelley et al., 2015) et HHPred (Steinegger et al., 2019a) permettent de rechercher des similitudes entre la séquence à modéliser et celles présentes dans les banques de données de structures 3D (e.g. PDB, CATH, SCOPe). Le programme MODELLER (Webb et Sali, 2021), référence du domaine de la modélisation comparative, permet de construire des modèles de structures 3D satisfaisant les contraintes spatiales sur la base de l'alignement d'une séquence avec une ou plusieurs séquences de structure 3D connue(s) et utilisée(s) comme gabarit(s), avant de les affiner par minimisation énergétique (étapes détaillées Figure 1.10). Des outils comme SWISS-MODEL (Waterhouse et al., 2018) permettent d'automatiser l'ensemble de ces étapes, depuis la recherche de gabarits jusqu'à la proposition de modèles de structures 3D affinés. Ces approches reposent donc très fortement sur les méthodes décrites dans le chapitre 1.1.4.2. Elles ne sont performantes que dans les cas où on retrouve de la similitude dans les banques de données structurales. Une alternative est proposée par les méthodes d'enfilage (« threading »), telles ORION (Ghouzam et al., 2015) et THREADER (Buchan et Jones, 2017), qui permettent de d'aller au-delà de la seule information évolutive en considérant des informations relatives aux caractéristiques structurales des séquences comparées.

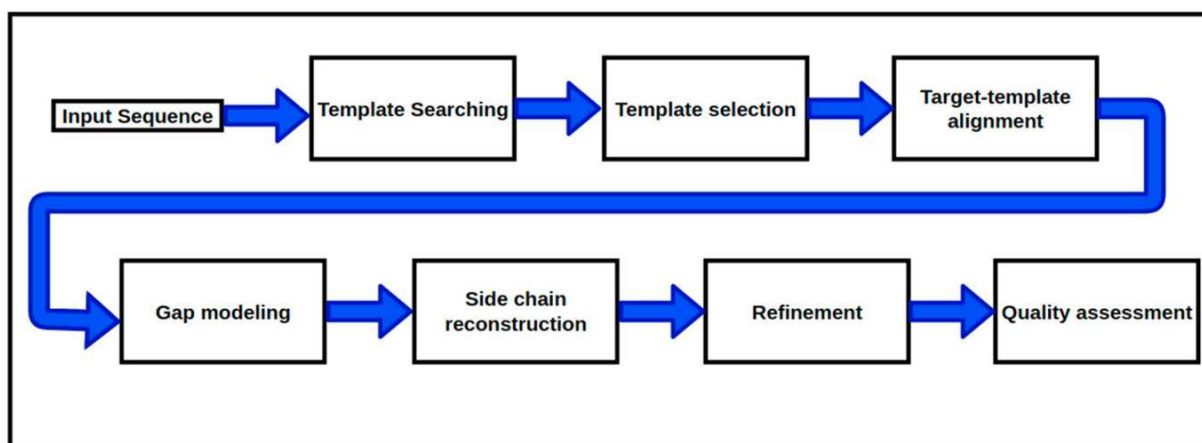


Figure 1.10. Principe de la modélisation basée sur l'homologie. Figure extraite de Jisna & Jayaraj (2021).

Depuis quelques années, les méthodes d'apprentissage profond basées sur l'information évolutive sont reconnues comme étant les plus efficaces pour les prédictions de structures 3D (Kandathil et al., 2019; Laine et al., 2021). Plusieurs prédicteurs de ce type peuvent être cités : DMPFold (Greener et al., 2019), RoseTTAFold (Baek et al., 2021), ou encore AlphaFold2 (Jumper et al., 2021). Ci-dessous je présenterai AlphaFold2, dont j'ai utilisé les prédictions au cours de mon travail de thèse.

AlphaFold2 (AF2), développé par DeepMind<sup>5</sup>, a révélé lors du concours CASP14 des niveaux d'efficacité inédits, comme on peut l'observer en Figure 1.11 (Pereira et al., 2021).

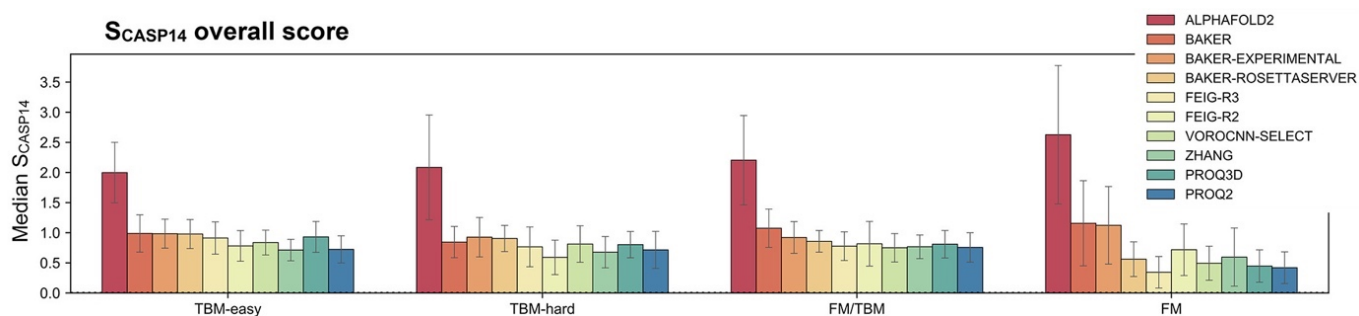


Figure 1.11. Barplots représentant le classement des 10 méthodes ayant obtenu les meilleurs scores (en ordonnées) lors du concours de prédiction CASP14. Chaque catégorie notée sur l'axe des abscisses correspond au niveau de difficulté des séquences à prédire. TBM signifie *Template Based Model*, soit modélisation basée sur un gabarit. FM signifie *Free Modelling*, indiquant qu'il n'y avait pas d'homologues dans les banques de données de structures 3D. Figure extraite de Pereira et al. (2021).

<sup>5</sup> <https://www.deepmind.com/>



AlphaFold2 (Jumper et al., 2021; Tunyasuvunakool et al., 2021) utilise des réseaux de neurones profonds pour prédire la structure des protéines à partir de leur séquence en acides aminés, son fonctionnement est schématisé en figure 1.12. Deux modes de représentations sont tirés de la séquence : (i) un alignement multiple, construit après recherche de la séquence dans diverses banques de données de séquences (BFD, MGnify, PDB, UniRef et UniProt), l'identification de motifs de co-variation (co-évolution) dans cette représentation permet de prédire les contacts entre chaque paires de résidus ; et (ii) une matrice de distance des résidus deux à deux, qui peut être complétée par l'utilisation de l'information structurale lorsque celle-ci est disponible. Ces modes de représentation sont utilisés par les deux modules composant AF2 pour prédire la structure de la protéine. De façon simplifiée, le module *evoformer* permet d'affiner la matrice de distances et l'alignement multiple des séquences en les confrontant l'un à l'autre. Ces deux représentations sont ensuite utilisées par le module *structure*, pour construire le modèle tridimensionnel de la protéine.

Le modèle de structure 3D ainsi obtenu est évalué par un score calculé par résidu, le « predicted local-distance difference test » (pLDDT), qui est une prédiction du score  $C\alpha$ -lddt (Mariani et al., 2013), développé pour évaluer l'efficacité des prédicteurs de structure en comparant les structures protéiques prédites aux structures déterminées expérimentalement. Il est obtenu en calculant les distances entre les  $C\alpha$  des résidus voisins dans la structure expérimentale d'une part, et dans la structure prédite d'autre part, puis en comparant ces valeurs le ratio de distances conservées dans la structure prédite par rapport à la structure expérimentale est calculé, c'est le  $C\alpha$ -lddt. La prédiction de cette valeur par AF2 est intégrée au réseau de neurones du module *structure* qui a été entraîné sur des structures PDB de bonne qualité (résolution entre 0.1 et 3.0 Å, pas de structure RMN) (Jumper et al., 2021). Un pLDDT supérieur 90 correspond à une prédiction de très bonne qualité, un pLDDT compris en 70 et 90 de bonne qualité, entre 50 et 70 de qualité faible, et inférieur à 50 de qualité très faible. Un exemple de prédiction 3D colorée en fonction du pLDDT est représenté Figure 1.13.

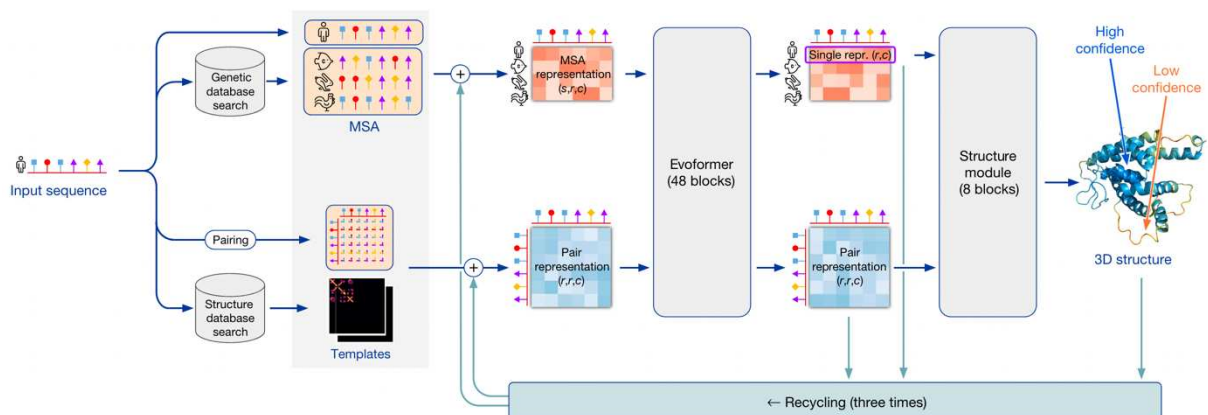


Figure 1.12. Pipeline d'AlphaFold2. Image extraite de Jumper et al. (2021).

Un second score est proposé par AF2, indépendamment de la structure 3D prédite : l'erreur d'alignement prédite (PAE pour *Predicted Alignment Error*). Le PAE est calculé par paires de résidus, il reflète le degré de confiance que donne AF2 à la position relative des deux résidus formant la paire. Un PAE faible indique que les orientations relatives sont bien prédites, au contraire, un PAE élevé indique qu'elles sont incertaines. Des PAE élevés sont observés au sein des domaines, et permettent de juger de la qualité des interactions inter-domaines.

AF2 a été appliqué à un ensemble de protéomes de référence et de protéomes présentant un intérêt pour la recherche médicale, puis étendu à l'ensemble des protéines répertoriées dans la banque de données UniProt. Ces prédictions sont répertoriées dans une banque de données dédiée (*AlphaFold Protein Structure DataBase* (AFDB) (Varadi et al., 2022), maintenant également accessible depuis les entrées UniProt.

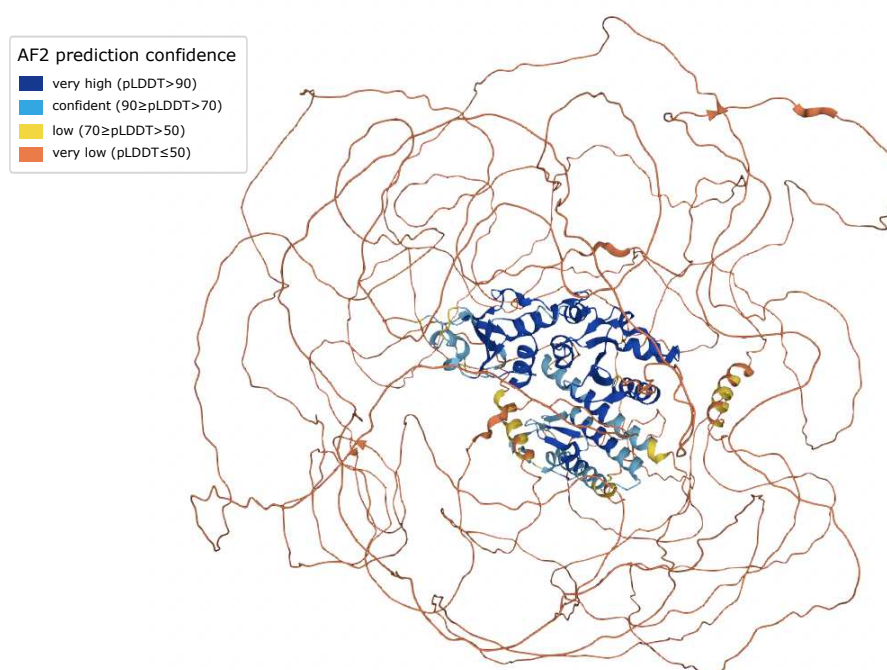


Figure 1.13. Structure prédite par AlphaFold2 pour BRAC1 humaine, colorée en fonction du pLDDT.

AF2 est très efficace, y compris pour des séquences qui n'ont pas d'homologues dans les banques de données structurales (scores pour la catégorie « FM » Figure 1.11) (Pereira et al., 2021). La comparaison des modèles produits par AF2 et des structures déterminées expérimentalement a déjà permis de révéler des homologies lointaines, indétectables par comparaison de séquence (e.g. Monzon et al., 2022). Cependant, il est vraisemblable que l'outil sera limité pour les séquences pour lesquelles il ne retrouve pas, ou très peu, d'homologues dans les banques de données de séquences. En effet, sans pouvoir déterminer les distances à partir des informations de co-variations de résidus, la qualité de prédiction devrait être

nécessairement moins bonne. D'autre part, AlphaFold2, comme tous les autres prédicteurs utilisant des algorithmes de *machine learning*, ne permet pas d'améliorer notre compréhension des principes du repliement des protéines (Outeiral et al., 2022), ce qui limite en particulier sa portée pour la prédiction de structures incluant des motifs de repliement non encore décrits à ce jour.

En l'absence d'homologues, il est donc important de pouvoir disposer d'outils permettant d'apprécier le potentiel de repliement (ou foldabilité) d'une séquence et d'en caractériser les éléments constitutifs. Dans le chapitre suivant, je présente les principes d'une approche de ce type, mise en place au laboratoire, que j'ai plus particulièrement utilisée dans ce travail de thèse.

#### ***1.1.4.4. Identification de signatures du repliement : Hydrophobic Cluster Analysis (HCA)***

Contrastant avec les approches précédemment présentées, la méthode Hydrophobic Cluster Analysis (HCA) exploite la dichotomie hydrophobe/hydrophile propre à un milieu micellaire au sein d'une représentation bidimensionnelle. Cette approche permet de révéler des informations relatives aux structures secondaires régulières, châssis de l'architecture des protéines. En particulier, HCA donne accès aux acides aminés constitutifs du cœur des structures de domaines globulaires et participant aux structures secondaires régulières, de façon intrinsèque (sans apprentissage) et à partir de l'analyse d'une séquence unique (sans se référer à l'ensemble des séquences d'une famille). Ceci revêt donc un intérêt majeur dès lors que l'on a affaire à des séquences orphelines, sans homologue connu. Après en avoir décrit les principes, je décrirai comment cette approche peut être utilisée pour décrypter différentes facettes propres aux caractéristiques structurales des protéines.

##### **A) Principes de la méthode HCA**

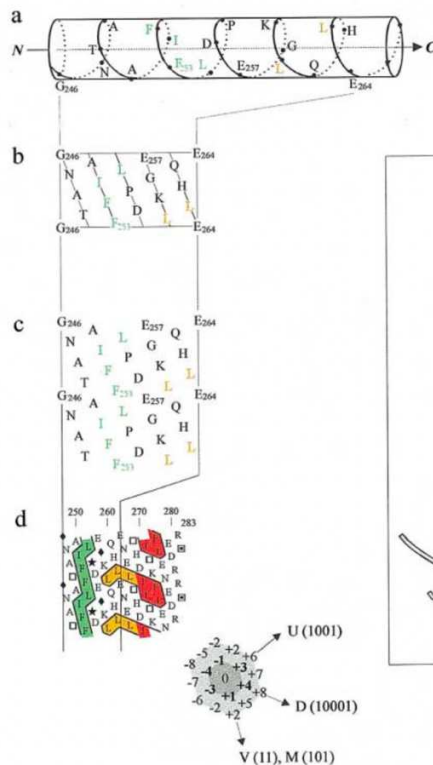
HCA permet de mettre en évidence la position des structures secondaires régulières des protéines via l'utilisation d'une représentation bidimensionnelle de leurs séquences dans le référentiel de l'hélice alpha (Gaboriaud et al., 1987; Woodcock et al., 1992). Comme illustré en Figure 1.14, le passage d'une séquence protéique 1D à la représentation 2D (d) est réalisé en positionnant les résidus sur la trame d'une hélice alpha (a), puis en coupant l'hélice le long de l'axe horizontal (b), permettant ainsi de passer de la tridimensionnalité de l'hélice à une représentation bidimensionnelle, où chaque ligne d'acides aminés correspond à un tour d'hélice. Ce plan est dupliqué (c), conduisant à un diagramme HCA bidimensionnel (d). Pour faciliter la lecture de ces diagrammes et faire apparaître au mieux des traits structuraux marquants, certains résidus sont remplacés par des symboles, et les résidus hydrophobes forts sont entourés, mettant en évidence les amas hydrophobes.

human  $\alpha$ 1 antitrypsin

1D

246 ...GNATAIFFLPDEGKLOHENEETHDIIITKFLNEDRRS... 283  
 ...♦NA□AIFFL★DEGKLOHENE□HDII□KFLNEDRR□...  
 ...00000111100000100100010001100110000000...

2D



3D

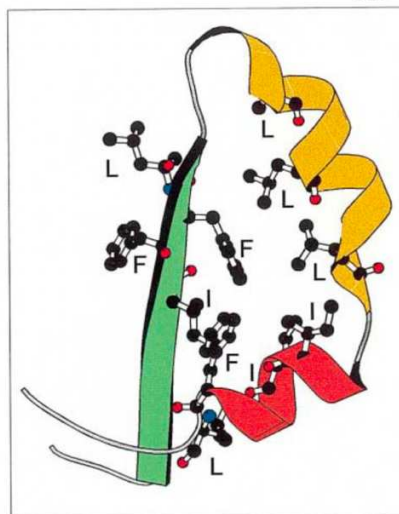


Figure 1.14. Représentation HCA : de la séquence (1D) au diagramme 2D et correspondance avec la structure 3D. Les acides aminés hydrophobes sont représentés en couleur, participant aux faces internes des structures secondaires régulières (brin beta interne – vert – et hélice alpha amphipatique et coudée – orange et rouge). Le code binaire (1 = V, I, L, M, F, Y, W ; 0 = autres acides aminés) est représenté sous la séquence 1D, ainsi que les symboles utilisés pour représenter les résidus glycine, proline, thréonine et sérine. Le voisinage 2D d'un résidu (noté 0) ainsi que les trois axes associés à la représentation bidimensionnelle sont également indiqués, avec les quatre amas basiques (C, M, U, D) à partir desquels l'ensemble des amas peuvent être construits (Rebehmed et al., 2016).

La définition des amas hydrophobes, détaillée ci-dessous, est également illustrée Figure 1.14-1D.

**L'alphabet HCA.** La définition des amas (ou clusters) hydrophobes repose sur l'utilisation d'un alphabet incluant 7 résidus hydrophobes forts. Ces résidus exercent un rôle moteur dans le repliement des protéines, comme indiqué par leur propension élevée à être présents dans des structures secondaires régulières (Callebaut et al., 1997; Hennetin et al., 2003) (Figure 1.15). Cet alphabet optimal, conduisant à observer un maximum de correspondance entre positions des structures secondaires régulières et des amas hydrophobes (Callebaut et al., 1997; Hennetin

et al., 2003; Soyer et al., 2000; Woodcock et al., 1992), comprend : la valine (V), l'isoleucine (I), la leucine (L), la phénylalanine (F), la méthionine (M), la tyrosine (Y) et le tryptophane (W). Parmi les résidus qui ne sont pas intégrés à cet alphabet, il est intéressant de remarquer que dans certaines conditions, la thréonine (T) et la sérine (S) (symbolisées respectivement par un carré avec et sans point sur le tracé HCA, et capables de former des liaisons hydrogène avec les atomes de la chaîne principale) ainsi que l'arginine (R), la lysine (K), possédant toutes deux une longue chaîne aliphatique, peuvent remplacer des résidus hydrophobes dans les structures secondaires. L'alanine (A) a, quant à elle, une propension très forte pour les hélices alpha. La cystéine (C), qui peut être retrouvée dans des milieux polaires, peut également jouer un rôle d'une importance similaire aux résidus hydrophobes dans le repliement des protéines par la formation de ponts disulfures, son exclusion de l'alphabet HCA est donc sujet à débat. Lorsque l'approche HCA est automatisée, celle-ci sera intégrée à l'alphabet, comme nous le verrons plus loin. D'autre part, une importance particulière est également donnée à la glycine (G) (symbolisée par un losange) et la proline (P) (symbolisée par une étoile), que l'on retrouve souvent dans des boucles (Figure 1.15).

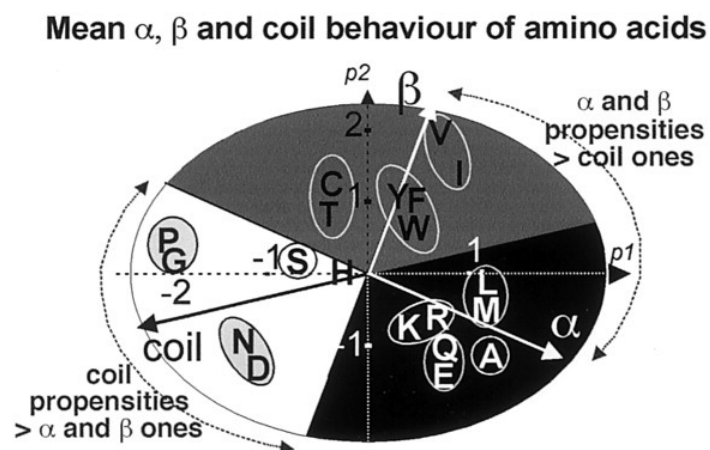


Figure 1.15. Représentation en 2D de la propension des acides aminés pour des états alpha, beta et coils. Les résidus retrouvés dans la zone blanche ont une propension plus importante pour les boucles (coils), ceux retrouvés dans la zone grise pour les brins beta, et ceux dans la zone noire pour les hélices alpha. Figure extraite de Hennetin et al. (2003).

**Distance de connectivité.** C'est la distance maximale (en nombre d'acides aminés) pour laquelle deux résidus sont en contact direct au sein du tracé HCA (Figure 1.14). La distance de connectivité couramment utilisée est celle observée dans une hélice alpha, soit une distance de 4 résidus. En effet, sur la trame d'une hélice alpha (caractérisée par un pas de 3.6 résidus) un résidu en position  $i$  sera en contact avec les positions  $i-1$ ,  $i-3$ ,  $i-4$  et  $i+1$ ,  $i+3$ ,  $i+4$  (Figure 1.14). Pour prendre en compte la périodicité des acides aminés hydrophobes propre aux brins beta, on considère également les connectivités à  $i-2$  et  $i+2$  (Figure 1.14). Il a été montré que la distance de connectivité de 4 est celle permettant une meilleure concordance entre structures secondaires et amas hydrophobes (Callebaut et al., 1997; Woodcock et al., 1992). Ainsi, deux résidus

hydrophobes forts appartiennent à des amas distincts s'ils sont séparés par au moins 4 autres résidus, ou par une proline (considérée comme rupteur).

**Codification des amas.** Pour faciliter la détection automatique d'amas et mettre en évidence la présence de motifs portés par les acides aminés hydrophobes au sein des séquences, une séquence binaire est générée à partir de la séquence en acides aminés, dans laquelle les résidus de l'alphabet HCA sont remplacés par des '1', et les autres résidus, hormis la proline, par des '0'. Un amas est alors défini par une suite de résidu commençant et terminant par un '1', ne contenant pas plus de quatre '0' à la suite, et ne contenant pas de proline (Figure 1.14).

À partir du code binaire, on peut également définir un code Peitsch (code décimal), un code plus compact et donc utile pour les applications informatiques (optimisation de mémoire). Ce code est calculé selon cette formule :

$$\text{code Peitsch} = \sum_{i=1}^n \text{bin}_i * 2^{n-i}$$

où  $i$  est la position du résidu allant de 1 à  $n$ ,  $n$  étant la taille de la séquence et  $\text{bin}_x$  est le code binaire trouvé en position  $x$  (on utilisera '0' pour la proline). Par exemple, pour le cluster « WWERY » le code binaire est « 11001 » le code Peitsch est 25 ( $1 \times 2^4 + 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0$ ).

## B) HCA, révélateur des caractéristiques du repliement des protéines

### (i) Les amas hydrophobes, briques conservées du repliement protéique

Il a été montré que les amas de taille supérieure à 2 résidus correspondent majoritairement à des structures secondaires régulières (Hennetin et al., 2003; Woodcock et al., 1992). Les amas portent une information plus différenciée vis-à-vis des structures secondaires régulières que les motifs binaires simples, dont ils se distinguent via la distance de connectivité considérée pour les construire (Hennetin et al., 2003). Par ailleurs, en accord avec la périodicité des acides aminés hydrophobes au sein des deux types de structures secondaires régulières, certains amas ont des affinités marquées soit pour des hélices alpha, soit pour des brins beta (Eudes et al., 2007; Rebehmed et al., 2016).

Ces affinités ont été calculées pour des protéines globulaires solubles dont les structures ont été déterminées expérimentalement (séquences et structures recensées dans la SCOPe), permettant ainsi de constituer **HCDB** (v1 : Eudes et al., 2007; v2 : Lamiable et al., 2019), une banque de données associant chaque cluster à une structure secondaire préférentielle, reprise sous le terme d'affinité (H/h et E/e pour des affinités forte/faible respectivement pour les hélices et les brins) (Figure 1.16). Sur 476 clusters recensés dans HCDB (clusters retrouvés avec une occurrence supérieure à 30 ce qui représente 82.7% du nombre total de clusters), 108 sont de la catégorie H, 84 E, 117 h et 135 e. Ces affinités ne prennent en compte que l'information binaire, dont il a été montré qu'elle prévaut sur l'information de séquence (Rebehmed et al., 2016) :

l'analyse peut être cependant affinée en prenant en compte les profils de séquences spécifiques des états majoritaires (vis-à-vis desquels les affinités sont déterminées) et minoritaires (Lamiable et al., 2019).

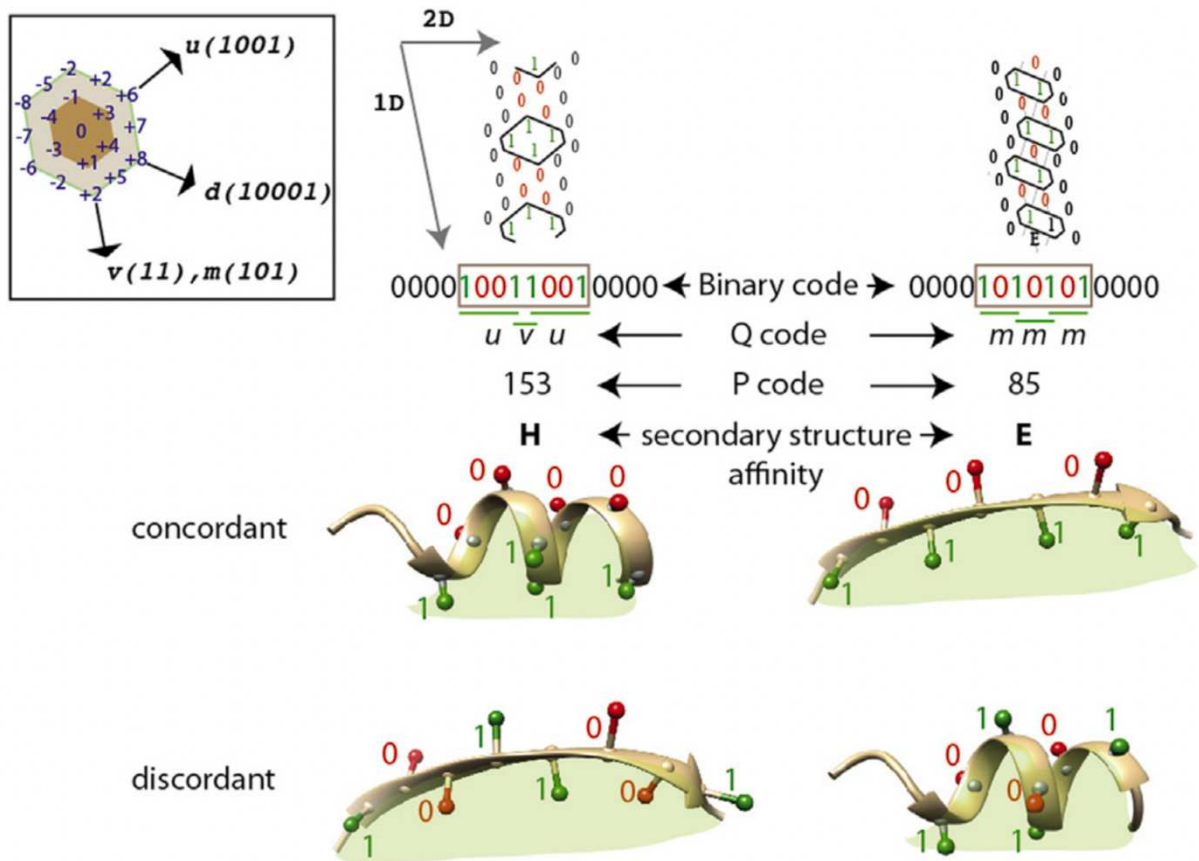


Figure 1.16. Préférence des amas hydrophobes pour les structures secondaires.

Les amas sont représentés par leur code binaire, Quark (Q) et Peitsch (P). Dans le code Q (également défini sur le réseau 2D illustré à gauche),  $v(11)$  signifie "vertical",  $m(101)$  signifie "mosaïque",  $u(1001)$  signifie "haut" et  $d(10001)$  signifie "bas". Les deux exemples montrés sont des amas fortement associés aux hélices alpha (P-153) et aux brins beta (P-85), selon HCBd v2. Dans les amas concordants (c'est-à-dire les amas pour lesquels l'état observé est en accord avec le comportement attendu du modèle binaire), toutes les positions hydrophobes (surlignées en vert sur la représentation 3D) sont enfouies dans le noyau hydrophobe (vert ombré). Dans les amas discordants (minoritaires), certaines des positions hydrophobes sont exposées au solvant. Figure extraite de Lamiable et al. (2019).

L'analyse des amas hydrophobes au sein de familles de séquences apparentées a montré que le caractère hydrophobe d'environ la moitié des positions est conservé, conduisant à identifier des signatures de repliement stables, même à très haut niveau de divergence évolutive (Fourty et al., 2008; Poupon et Mornon, 1998). Ainsi, des schémas de substitution préférentiels des amas hydrophobes peuvent être mis en évidence (Leduc, 2006). Ces signatures de repliement, restant stables en dépit d'un grand nombre de variations : substitutions mais aussi insertions/délétions d'acides aminés, peuvent être exploitées afin de mettre en évidence des

homologies lointaines et éventuellement, à identifier de nouvelles familles de domaines (e.g. Callebaut et al., 1999, 2001, 2002; Callebaut & Mornon, 1997, 2010).

D'un point de vue fondamental, la comparaison des caractéristiques des amas hydrophobes dans des ensembles de séquences réelles et de séquences aléatoires a permis d'identifier les briques préférentiellement utilisées par la nature pour construire les repliements 3D de protéines (solubles mais également membranaires) et évaluer leur robustesse par rapport à l'évolution (Figure 1.17) (Callebaut et al., 1997).

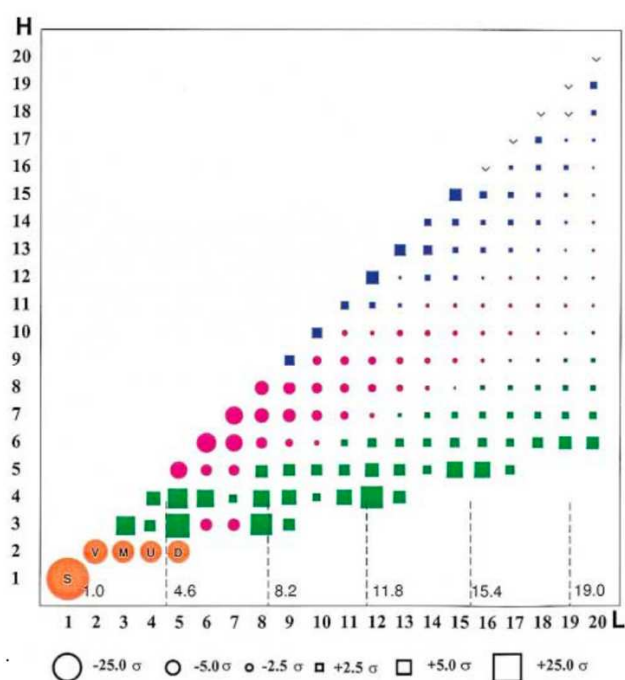


Figure 1.17. Représentation du comportement moyen des amas hydrophobes. L est la taille de l'amas hydrophobe (en nombre d'acides aminés), H est le nombre d'acides aminés hydrophobes. Les carrés indiquent les amas préférentiellement adoptés par les protéines (Z-scores positifs), leur surface correspondant au Z-score moyen. Les cercles indiquent les amas relativement écartés (Z-scores négatifs). La diagonale est occupée par des amas composés uniquement de résidus hydrophobes. Les autres positions occupées ont un nombre variable d'amas « isotopes » (mêmes valeurs de L et H). Par exemple, en position L=8 et H=4 il y a 12 isotopes possibles. Le nombre d'amas possible augmente rapidement avec la taille de l'amas et le nombre d'acides aminés (il y en a, par exemple, 216 pour L=12 et H=7). Dans cette représentation environ  $10^6$  amas sont représentés. On peut remarquer (i) la faible représentation des petits amas S (1), V (11), M (101), U (1001), D (10001) (en bas à gauche), (ii) une large zone favorable (carrés verts), renforcée au niveau de la périodicité de l'hélice alpha (4.6, 8.2, 11.8, 15.4, 19.0 acides aminés) et correspondant aux briques préférées par les domaines globulaires, (iii) une région non favorable (cercles violets) puis enfin iv) une dernière région favorable (carrés bleus) correspondant à des segments transmembranaires (e.g. 12 acides aminés hydrophobes consécutifs). Figure extraite de Callebaut et al. (1997).



## (ii) Amas hydrophobes et segments foldables

L'analyse de la « texture » en amas hydrophobes des protéines permet de mettre en évidence des régions homogènes en termes de composition en amas hydrophobes et correspondant à des segments de séquences qui ont la capacité de se replier dans un environnement soluble ou membranaire, par la suite nous les nommerons segments « foldables ». Ces segments, faciles à repérer sur les tracés HCA (encadrés dans la Figure 1.18), peuvent également être définis automatiquement grâce à l'outil segmentation-HCA (seg-HCA) (Faure et Callebaut, 2013).

La définition automatique des segments foldables est réalisée selon les étapes suivantes :

- 1- Identification des amas hydrophobes, en incluant la cystéine (C) à l'alphabet HCA (VILMFYW). Pour la suite, seuls les amas de taille supérieure à 2 résidus seront pris en compte (Figure 1.18.A).
- 2- Création d'une séquence protéique binaire (qui ne correspond pas aux codes binaires des amas vus précédemment), où un résidu dans un amas (qu'il soit hydrophobe ou pas) sera noté '1', un résidu hors amas sera noté '0' (acides aminés polaires ou hydrophobes isolés, en rouge sur la Figure 1.18.A).
- 3- Calcul du pourcentage de résidus au sein d'amas sur une fenêtre glissante de 17 résidus (représenté en Figure 1.18.B).
- 4- Identification des régions denses en amas hydrophobes (zones avec un pourcentage de résidus en clusters > 10%), typiques des régions foldables (limite rouge Figure 1.18.B).
- 5- Construction d'un arbre où les feuilles sont les amas hydrophobes et où les distances sont le nombre de résidus séparant les amas.
- 6- seg-HCA compare les différentes régions définies, d'une part, par les nœuds dans l'arbre (annotées 1 dans l'arbre) et d'autre part, par la limite fixe de 10% de résidus en amas, et retient le découpage permettant le meilleur recouplement entre les deux (annoté 3 dans la figure).

Un outil, pyHCA, a été développé au laboratoire par Tristan Bitard-Feildel pour adapter la méthode seg-HCA en langage python sous sa commande *segment* (également utilisable en ligne de commande : *hcatk segment*).

## 1.1. L'ordre dans les protéines : relations séquence-structure-fonction

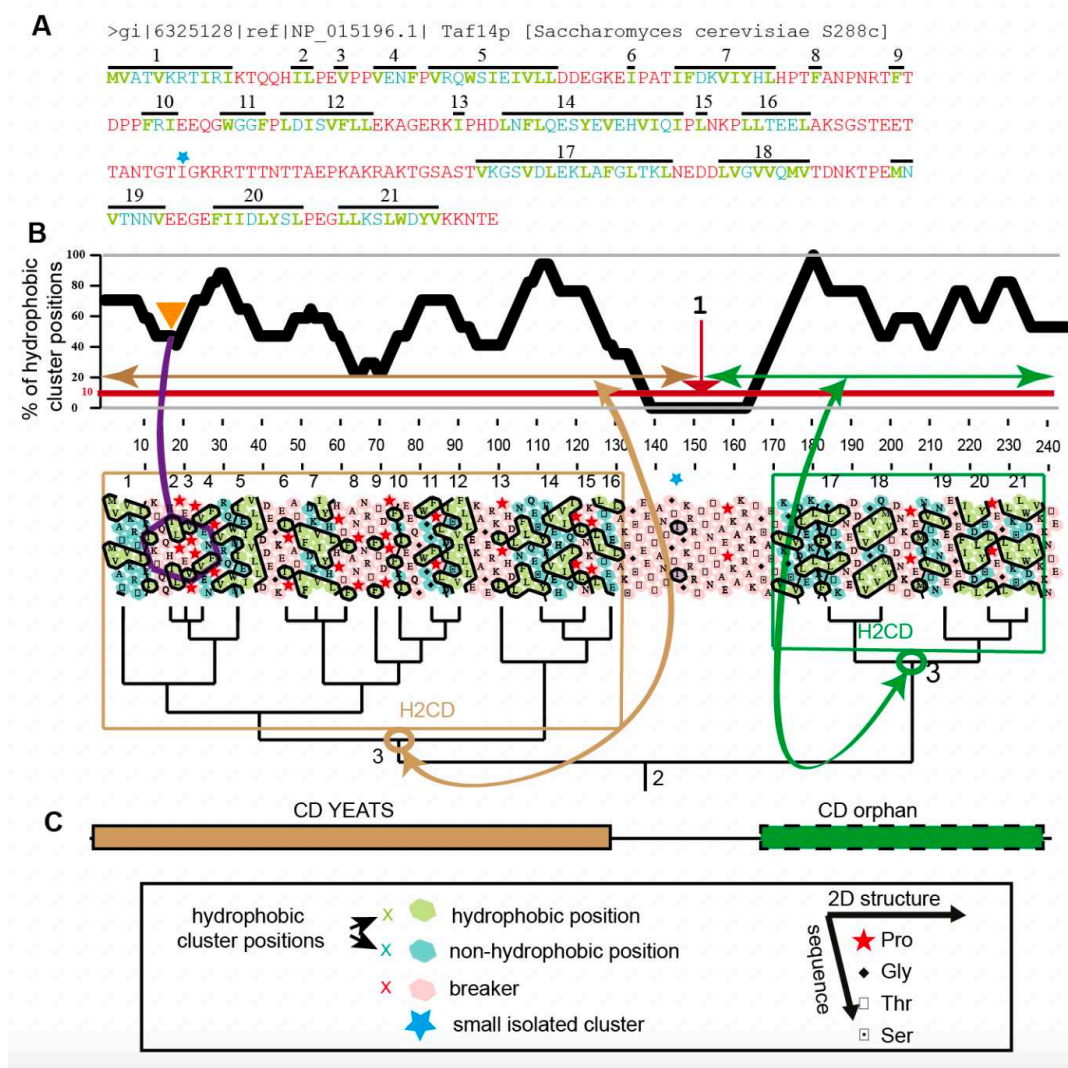


Figure 1.18. Principe de seg-HCA. A) Séquence protéique : acides aminés (aa) hydrophobes forts en vert, aa non hydrophobes dans des amas hydrophobes en bleu, aa dans les segments inter-amas en rouge. Les amas hydrophobes sont mis en évidence par un trait noir. B) Graphe représentant le pourcentage d'aa au sein des amas hydrophobes sur des fenêtres glissantes de 17 aa, en vis-à-vis du tracé HCA de la séquence, et de l'arbre des amas hydrophobes. C) seg-HCA identifie 2 segments foldables dans la séquence, le premier schématisé en jaune et le second en vert. CD = domaine conservé, le premier correspondant à un domaine YEATS. Figure extraite de Faure & Callebaut (2013).

La validité de ce découpage a été appuyée par confrontation des prédictions réalisées sur cinq protéomes de référence (*H. sapiens*, *S. cerevisiae*, *P. falciparum*, *E. coli*, *A. fulgidus*) avec des banques de structures expérimentales (PDB et SCOP classes *a* à *f* (domaines solubles et membranaires) et avec des banques de domaines (NCBI Conserved Domain Database (CDD)) (Faure et Callebaut, 2013). Les séquences de ces banques sont bien couvertes par les segments foldables (97.8 % et 97.5 % pour les banques PDB et SCOPe, 75 % pour la banque CDD). Les exceptions concernent des domaines dont le moteur de repliement ne fait pas intervenir d'acides aminés hydrophobes, organisés en cœur compact (certains domaines de liaison d'ions).

## 1.2. Les régions intrinsèquement désordonnées

Les régions intrinsèquement désordonnées (ou IDRs pour *intrinsically disordered regions*) sont des régions qui, au sein des protéines, n'adoptent pas de structure 3D bien définie et stable, du moins, pour une partie d'entre elles, pas de manière spontanée (Habchi et al., 2014; Tompa, 2002; Uversky et al., 2000; van der Lee et al., 2014; Wright et Dyson, 1999). Lorsque ces régions couvrent l'intégralité de la protéine, on parle de protéines désordonnées, ou IDPs pour *intrinsically disordered proteins*. Le caractère désordonné de ces séquences en acides aminés est généralement associé à une pression de sélection moindre ; les séquences évoluent rapidement, et il est généralement difficile de mettre en avant des relations d'homologie (Zarin et al., 2019). L'annotation de ces séquences est aussi critique, car leur caractère désordonné ne signifie pas absence de fonction. Au contraire, leur flexibilité et caractère dynamique leur confère des propriétés largement exploitées dans divers processus cellulaires, dont l'importance a été largement démontrée au cours des vingt dernières années (van der Lee et al., 2014). De plus, les IDRs sont très répandus dans le vivant, il est admis qu'ils couvrent environ 30% des résidus du protéome humain, et jusqu'à 50% dans certains protéomes d'eucaryotes unicellulaires. Chez les procaryotes, la part de désordre est plus faible, souvent inférieure à 28% (Peng et al., 2015; Ward et al., 2004; Xue et al., 2010).

### 1.2.1. IDRs : une hétérogénéité de conformations

Une région protéique complètement désordonnée ne présente aucune structure secondaire régulière et la chaîne polypeptidique est complètement étendue. Mais entre cet extrême et une chaîne polypeptidique repliée et compacte typique des domaines globulaires, plusieurs niveaux de repliement sont possibles (Figure 1.19). Quatre états intermédiaires de repliement ont ainsi été décrits (van der Lee et al., 2014) : les conformations étendues (*coils* ou *Extended*), caractérisées par l'absence totale de structure organisée (on les retrouvera aussi sous le nom de « désordre complet »); les globules pré-fondus (*pre-molten globules*), dans lesquels apparaissent des structures secondaires transitoires (regroupant les catégories « *transient secondary structure* » et « *compact globule* »); les globules fondus (*molten globules*), qui présentent des structures secondaires semblables à celles d'un état natif, et une structuration partiellement repliée et stable ; et enfin, l'*ordre* (dans lequel on retrouve les catégories « *disordered loop* » et « *folded protein*») où le désordre correspond aux boucles et aux régions de liaison (*linkers*), dans le cas des protéines multi-domaines, de taille plus ou moins importante (Dunker et al., 2001). Les IDRs sont caractérisés par une absence de contrainte structurale à l'état isolé, ce qui leur confère une grande flexibilité et un caractère dynamique important. Ainsi, elles pourront osciller rapidement entre différentes conformations qui sont conditionnées à leur environnement (e.g. pH, température) ou à leur liaison à un partenaire (Dyson et Wright, 1998; Uversky et Dunker, 2010).

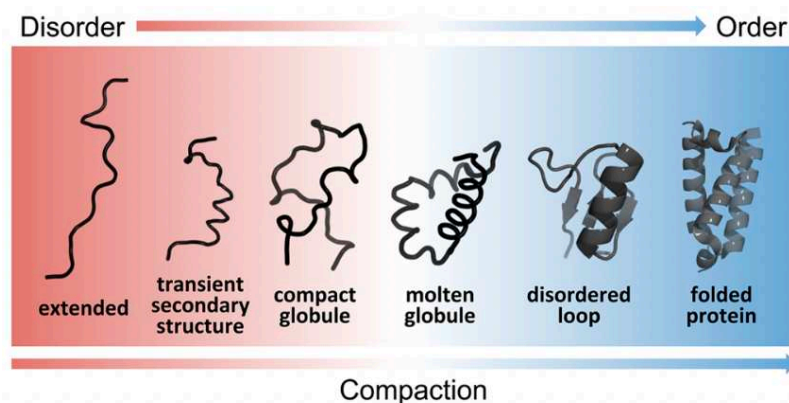


Figure 1.19. Structures protéiques 3D situées à différents niveaux sur le continuum désordre/ordre. Figure extraite de van der Lee et al. (2014).

## 1.2.2. La flexibilité des IDRs : un avantage fonctionnel

### 1.2.2.1. Les fonctions des IDRs

Plusieurs fonctions sont décrites dans le désordre, et différentes classifications des IDRs basées sur ce critère ont été proposées (Dunker et al., 2001; Gsponer et Babu, 2009; van der Lee et al., 2014). Celle qui semble prédominer aujourd'hui est celle proposée par Tompa (2002, 2005) et représentée en Figure 1.20. Les IDRs qui ne peuvent pas interagir avec des partenaires et donc ne se structurent pas occupent la fonction de **chaîne entropique**. Ce sont souvent des régions de liaison (*linkers*), qui permettent le mouvement des domaines situés de part et d'autre de l'IDR, ou des « espaceurs » (*spacers*), qui régulent les interactions entre domaines. Les autres IDRs exercent généralement des rôles qui leur sont conférés par leur capacité à interagir avec des partenaires (protéines, ADN, ARN, petits ligands). Dans une grande partie des cas, cette interaction permettra à la région désordonnée de se structurer transitoirement (on parle de transition désordre-vers-ordre), ce qui lui conférera sa fonction (Jakob et al., 2014; Sugase et al., 2007; Wright et Dyson, 2009). En raison de leur flexibilité et leur caractère dynamique, les liaisons dans lesquelles sont impliquées les IDRs peuvent être d'affinité variable, et être ou non transitoires (Dreier et al., 2022; Wright et Dyson, 2015) et de spécificité variable, permettant à certains IDRs d'avoir plusieurs partenaires d'interaction (c'est par exemple le cas du suppresseur de tumeur p53 (Mohan et al., 2006; Oldfield et al., 2005)). Toutes ces particularités font des IDRs des points centraux (*hubs*) des réseaux de régulation ou de signalisation. Les IDRs pourront ainsi jouer un rôle de régulateurs par le biais de **sites de modifications** post-traductionnelles (*display sites*), comme on l'observe par exemple au niveau des queues d'histones (Kouzarides, 2007) ou du suppresseur de tumeur p53 (Bode et Dong, 2004). Les **effecteurs**, impliqués dans la signalisation cellulaire, sont capable d'activer ou d'inhiber certains acteurs protéiques (Galea et al., 2008; Sugase et al., 2007). Parmi les IDRs, on retrouve

également les **assembleurs** qui jouent un rôle de recrutement et de stabilisation dans la formation de complexes moléculaires, les **chaperonnes** qui assistent leur partenaire (protéine ou ARN) dans le processus de repliement (Schroeder et al., 2004; Young et al., 2004) et les **récepteurs éboueurs** (*scavengers*) qui stockent et neutralisent de petits ligands (tels que l'ATP ou l'adrénaline (Daniels et al., 1978)).

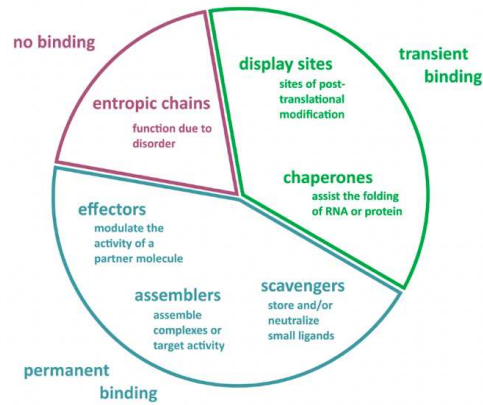


Figure 1.20. Classification fonctionnelle des IDR. Figure extraite de van der Lee et al. (2014).

Enfin, il a été montré que les IDPs peuvent jouer un rôle de séparateurs de phases liquide-liquide (LLPS pour « *Liquid liquid phase separation* »), créant ainsi des gouttelettes (*droplets*) au sein des cellules (Flock et al., 2014). Ces formations peuvent permettre la concentration de molécules dans un espace confiné de la cellule, favorisant des réactions biochimiques, ou au contraire, elles peuvent permettre d'isoler certaines molécules du reste de la cellule (Alberti, 2017; Cioce et Lamond, 2005). Elles sont également impliquées dans la signalisation cellulaire (Banani et al., 2017; Chong et Forman-Kay, 2016; Wu et Fuxreiter, 2016).

### 1.2.2.2. Les sites fonctionnels d'interaction

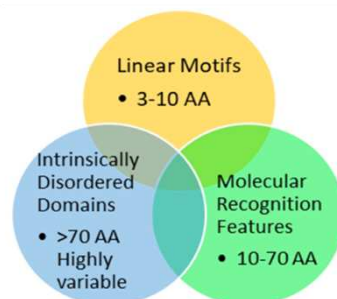


Figure 1.21. Trois catégories d'éléments d'interaction présents au sein des IDR. Figure extraite de Jones & Tepe (2019).

Les interactions des IDR avec leurs partenaires peuvent faire intervenir trois types de régions, présentées ci-dessous et en Figure 1.21. Il existe un recouvrement de ces caractéristiques, qui suggère que ce sont trois états situés sur un même continuum.

Les courts motifs linéaires (**Short Linear motifs (SLiMs)**) (Dinkel et al., 2014), aussi appelés motifs linéaires (*Linear Motifs* (LM) (Diella et al., 2008) et MiniMotifs (Mi et al., 2012) sont des motifs courts (3-10 acides aminés), qui sont à l'origine d'interactions de faible affinité (interactions transitoires et réversibles). Les SLiMs ne sont pas retrouvés systématiquement au sein de régions désordonnées, mais c'est le cas pour environ 80% d'entre eux (Davey et al., 2011; Fuxreiter et al., 2007). Parmi ces SLiMs, environ 60% évoluent d'un état désordonné à un état ordonné lors de l'interaction avec leur partenaire (Davey et al., 2011). Il est intéressant de noter ici que diverses études ont montré que certains IDR maintenaient un certain degré d'hétérogénéité conformationnelle lors de la liaison à un partenaire, amenant à intégrer une notion de flou (« fuzziness ») (Borgia et al., 2018; Sharma et al., 2015; Tompa et Fuxreiter, 2008). On distingue deux grandes familles de SLiMs, d'une part les sites de modifications post-traductionnelles, et d'autre part les ligands. Les SLiMs ligands permettent de recruter des molécules (protéiques ou nucléotidiques), jouant divers rôles. Le rôle d'assembleur, notamment, est facilité par la multivalence des SLiMs qui permet ainsi le recrutement de plusieurs partenaires de manière rapprochée. Un exemple est celui d'un motif intrinsèquement désordonné de liaison à l'ubiquitine (DisUBM), qui reste désordonné lors de sa liaison et agirait en augmentant l'affinité d'autres IDPs pour des partenaires repliés ubiquitylés (Dreier et al., 2022).

Les dispositifs de reconnaissance moléculaire (**Molecular Recognition Features (MoRFs)**) (Oldfield et al., 2005) sont des motifs plus longs (10-70 aa), spécifiques du désordre, aussi appelés éléments structuraux préformés (*performed structured elements* ou PSEs (Fuxreiter et al., 2004)) ou éléments de reconnaissance moléculaire (*molecular recognition elements* ou MOREs (Mohan et al., 2006; Oldfield et al., 2005)). Ces motifs se replient en interagissant avec leurs partenaires (la protéine p53, par exemple, contient plusieurs MoRFs qui sont en état désordonné en absence de partenaires (Oldfield et al., 2005)). Ils ont la particularité de présenter des structures secondaires préformées, qui se précisent et sont stabilisées lors de l'interaction. Comme les SLiMs, les MoRFs peuvent jouer le rôle d'assembleurs (Abet et al., 2014).

**Les domaines intrinsèquement désordonnés (IDDs)** ont la particularité d'être conservés, au même titre que les domaines structurés (Chen et al., 2006), mais ils sont partiellement ou complètement désordonnés (Tompa et al., 2009). C'est le cas, par exemple, du domaine inhibiteur de kinase (KID) des inhibiteurs de CdK (Galea et al., 2008). Ces domaines sont souvent impliqués dans des liaisons à l'ADN, l'ARN, ou à des protéines (Chen et al., 2006). D'autre part, certains domaines structurés fonctionnent uniquement en présence de désordre dans leur voisinage, les IDD avec lesquels ils co-existent sont donc conservés, bien qu'ils n'exercent pas de fonctions qui leur sont propres (Pentony et Jones, 2010; Teraguchi et al., 2010).

### 1.2.3. Les séquences du désordre : biais et évolution

La flexibilité des IDR est liée à leur composition en acides aminés. Des premières études (Tompa, 2002; Uversky et Dunker, 2010), étayées par des statistiques plus récentes (Figure 1.22 ; Chong & Mir, 2021), ont montré que leurs séquences sont enrichies en proline, en résidus polaires et chargés (notamment Glu), et appauvries en résidus hydrophobes forts (Phe, Trp et Tyr). La faible teneur en résidus hydrophobes de ces régions limite donc de manière très importante les interactions hydrophobes et rend impossible la formation d'un cœur hydrophobe stable en milieu soluble.

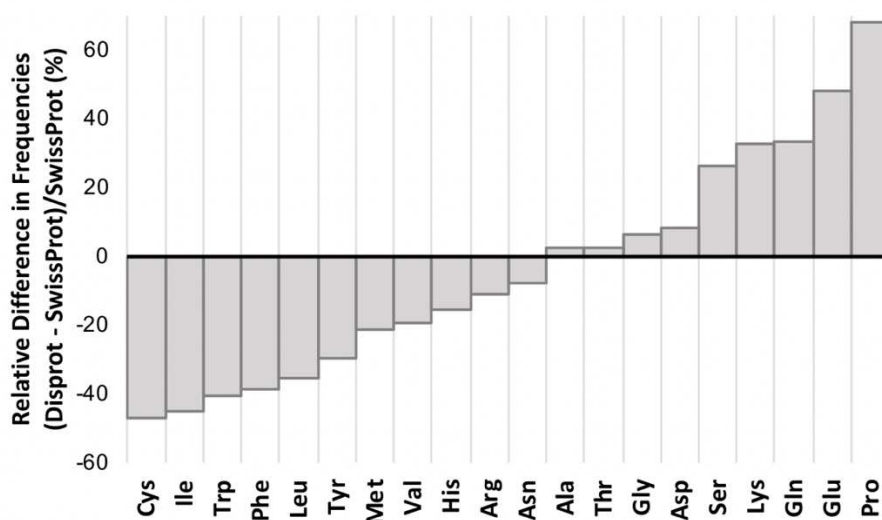


Figure 1.22. Biais compositionnel des séquences du désordre. Différence relative pour chaque acide aminé entre sa fréquence dans la banque DisProt (banque de séquences associées à des états désordonnés) et dans la banque SwissProt (banque de séquences). Les valeurs positives indiquent les résidus observés de manière plus importante dans le désordre. Figure extraite de Chong & Mir (2021).

Les IDRs ne rencontrent pas les mêmes contraintes que les domaines structurés, qui doivent maintenir des interactions fortes entre résidus pour maintenir l'intégrité de la structure. Les séquences désordonnées sont donc soumises à une pression de sélection moindre, et cela se reflète dans leur évolution plus rapide que celle des domaines structurés (Brown et al., 2011). Cependant, ces régions conservent leur caractère désordonné (Bellay et al., 2011), ce qui reflète leur importance dans les processus cellulaires.

Si les IDRs présentent des caractéristiques de séquences qui, globalement, les distinguent des protéines ordonnées, des précisions peuvent être apportées en s'intéressant à des motifs ou « *patterns* » spécifiques (SLiMs (Weatheritt et al., 2012), MoRFs (Vacic et al., 2007), régions de faible complexité (« Low complexity regions » (LCRs) / « compositionally biased regions » (CBRs): séquences composées de répétitions d'un ou quelques résidus) (Kato et al., 2012; Mier et al., 2020; Promponas et al., 2000; Wootton et Federhen, 1993). Une grammaire spécifique des motifs impliqués dans les séparations de phase liquide-liquide (LLPS) a également été proposée (Wang et al., 2018).

### 1.2.4. Les ressources et outils du désordre

Il existe plusieurs méthodes expérimentales permettant de déterminer si une protéine présente du désordre ou non, chacune présentant ses avantages. La spectroscopie à résonance magnétique (RMN) est la méthode de référence pour caractériser le désordre, elle fournit une information sur les ensembles conformationnels et la dynamique des protéines en solution (Dunker et al., 2001; Dyson et Wright, 2004). La cristallographie à rayons X permet d'identifier la position de résidus désordonnés indirectement, par absence de signal (Dunker et al., 2001). La diffusion des rayons X aux petits angles (SAXS (*small-angle X-ray scattering*)) permet d'accéder aux dimensions et à la forme de la protéine (Receveur-Bréchet et al., 2006). Le dichroïsme circulaire détecte l'absence de structure fixe dans les protéines (Chemes et al., 2012). Le transfert d'énergie par résonance Förster (FRET) mesure les dynamiques et conformations individuelles à l'état non lié (Schuler et Hofmann, 2013), ainsi que les frictions pendant le repliement. Enfin, la microscopie à force atomique (AFM) est utile à la caractérisation de l'hétérogénéité structurale de protéines isolées (Kodera et Ando, 2022).

#### 1.2.4.1. Les banques de données du désordre

Les informations sur les protéines et régions désordonnées qui ont été caractérisées expérimentalement sont recensées dans des banques de données du désordre, dont une revue a été récemment publiée (Piovesan et al., 2022). Cet article fournit une vue schématique de l'organisation de ces banques de données en fonction des différentes dimensions vis-à-vis desquelles les IDPs/IDRs peuvent être considérées (Figure 1.23), ainsi qu'une table comparative (Tableau 1.2). DisProt<sup>6</sup> (Hatos et al., 2020), base de donnée très largement utilisée et que nous avons plus particulièrement exploité dans le cadre de mes travaux de thèse, recense les séquences et fonctions d'IDRs/IDPs caractérisées expérimentalement. La qualité de cette banque repose sur le travail des curateurs, qui fournissent une annotation précise et exhaustive, régulièrement actualisée. Elle présente également l'avantage d'être couplée aux annotations d'UniProtKB et de présenter des classifications fines fonction du type de désordre (« structural (disordered) state », « structural transitions », « interaction partners », « fonctions »).

D'autres banques donnent accès à des informations spécifiques. Ainsi par exemple, FuZDB (Miskei et al., 2020) reprend les informations sur les régions floues (« fuzzy ») impliquées dans la formation de complexes et d'assemblages de complexité plus élevée. DIBS (Schad et al., 2018) et MFIB (Fichó et al., 2017) collectent des exemples de repliements induits lors de liaison, avec lien sur les structures expérimentales. La banque de données ELM (« eukaryotic linear motif ») collecte les informations relatives aux SLiMs. À noter que sont également reprises en Figure 1.23 des bases de données de prédiction (voir ci-après), dont celles obtenues via AlphaFold2. Ce panorama n'est cependant pas exhaustif, d'autres banques de données

---

<sup>6</sup> <http://www.disprot.org/>



n'étant pas mentionnées, dont par exemple celles reposant sur l'analyse d'alphabet structuraux (Akhila et al., 2020a, 2020b).

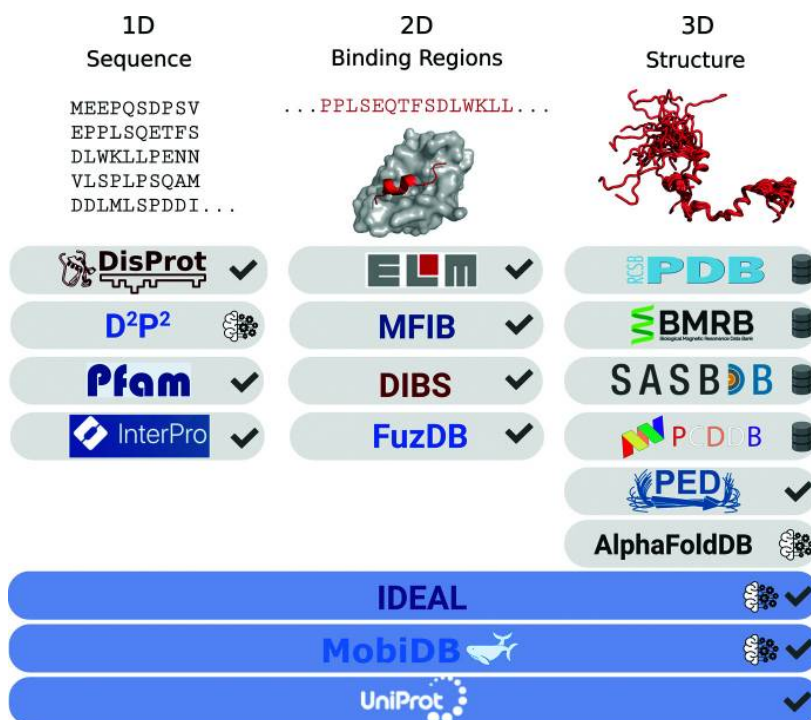


Figure 1.23. Les banques de données du désordre. Les icônes à côté des banques indiquent la nature de celles-ci : « marque de contrôle » = banques de données organisées, « database » = banques de dépôt, « machine-learning » = banques incluant des données de prédiction. Figure extraite de Piovesan et al. (2022).

Name	URL	Creation date	IDRs		LIPs	
			Proteins	Content (%)	Proteins	Content (%)
Pfam	<a href="http://pfam.xfam.org/">http://pfam.xfam.org/</a>	1997	39†	>80	—	—
UniProtKB	<a href="https://www.uniprot.org/">https://www.uniprot.org/</a>	2002‡	475‡	13.8	—	—
ELM	<a href="http://elm.eu.org/">http://elm.eu.org/</a>	2003	—	—	3542	1.4
DisProt	<a href="https://disprot.org/">https://disprot.org/</a>	2005	1746	20.5	729	19.3
IDEAL	<a href="https://www.ideal-db.org/">https://www.ideal-db.org/</a>	2012	995	10.3	317	8.9
FuzDB	<a href="https://fuzdb.org/">https://fuzdb.org/</a>	2016	110§	16.6§	—	—
MFIB	<a href="http://mfib.enzim.ttk.mta.hu/">http://mfib.enzim.ttk.mta.hu/</a>	2017	—	—	205	24.7
DIBS	<a href="http://dibs.enzim.ttk.mta.hu/">http://dibs.enzim.ttk.mta.hu/</a>	2017	—	—	772	4.1

Tableau 1.2. Banques de données du désordre. La date de création (« creation date ») correspond à la date de la première publication décrivant la ressource. La colonne « Proteins » indique le nombre de protéines avec IDRs ou SLiMs (notés ici LIPs), (données relevées en Octobre 2021). † familles des IDD disponibles dans Pfam, ‡ Protéines issues d'UniProtKB avec au moins une IDR annotée manuellement, § FuzDB annoté les IDRs qui forment des complexes flous (*fuzzy complexes*).

#### 1.2.4.2. Les prédicteurs du désordre

Les méthodes expérimentales sont couteuses et ne permettent pas de traiter des quantités importantes d'information. Dans ce contexte, de nombreux prédicteurs de désordre ont été proposés (Liu et al., 2019) depuis une bonne vingtaine d'années. Le premier prédicteur, PONDR (Romero et al., 1997), repose sur des réseaux de neurones utilisant des attributs de séquence sur des fenêtres glissantes de 9 à 21 résidus. Une initiative récente nommée CAID (*Critical Assessment of protein Intrinsic Disorder predictions* (Necci et al., 2021)) permet de confronter les nombreux prédicteurs à une même évaluation, et ainsi de comparer les performances de chacun. Cette expérience met en évidence quatre prédicteurs qui réalisent de bonnes prédictions en toutes situations : SPOT-Disorder2 (Hanson et al., 2019), RawMSA (Mirabello et Wallner, 2019), AUCpreD (Wang et al., 2020) et fIDPnn (Hu et al., 2021), ce dernier présentant la particularité de proposer une annotation fonctionnelle associée à la région désordonnée qu'il prédit). Ces quatre méthodes utilisent de l'**apprentissage profond** (*deep-learning*), ce qui leur confère notamment l'avantage de pouvoir prendre en compte à la fois les interactions à courte et à longue portée, un point limitant dans les autres prédicteurs qui doivent souvent traiter séparément les segments courts et longs (Liu et al., 2019). Elles reposent sur la considération de descripteurs de propriétés des séquences, d'informations structurales et d'informations évolutives. Ces méthodes sont donc de capacité plus limitée quand les séquences étudiées ne présentent pas de similitude avec d'autres protéines présentes dans les banques de données et qu'il n'y a donc pas d'information évolutive les concernant. En témoignent les différences de performance entre SPOT-disorder et sa variante SPOT-disorder-single (Hanson et al., 2018), qui s'affranchit de l'information évolutive (Necci et al., 2021).

En dehors des méthodes d'apprentissage automatique, trois autres classes de prédicteurs peuvent être décrites (Liu et al., 2019), parmi lesquelles on retrouve les **méthodes reposant sur des modèles**, très contraignantes car elles reposent sur la présence d'homologues dans des banques de données structurales, et les **méta-méthodes**, qui combinent plusieurs prédicteurs.

La dernière catégorie est celle des **méthodes reposant sur les caractéristiques physico-chimiques des séquences**, en particulier sur celles inhérentes aux propriétés du repliement des protéines. Par exemple, Globplot (Linding et al., 2003) calcule un score sur la séquence qui dépend de la propension des résidus à être dans des structures secondaires régulières ou en *random coil*. Uversky et al. (2000) montrent que la charge nette et l'hydrophobicité de la séquence sont de bonnes caractéristiques pour discriminer les IDPs. Enfin, d'autres prédicteurs reposent sur l'hypothèse que les résidus des IDRs ne parviennent pas à former suffisamment de connections pour atteindre une structure 3D stable (Dosztányi et al., 2005a; Galzitskaya et al., 2006; Schlessinger et al., 2007). C'est le cas de IUPred (Dosztányi et al., 2005b) et de sa version améliorée IUPred2 (Mészáros et al., 2018), qui estiment la capacité des séquences à former des contacts stabilisants lors du repliement. Pour cela, les caractéristiques physico-chimiques et l'interaction potentielle avec des partenaires de chaque résidu est prise en compte. Les paramètres utilisés par IUPred sont extraits d'un jeu de données de séquences protéiques de

référence (qui ne tient pas compte du désordre) (Dosztányi et al., 2005a). Ces prédicteurs ne sont pas les plus performants, mais ils ont l'avantage d'être très rapides tout en restant efficaces (Necci et al., 2021), ce qui permet de les appliquer à de grands jeux de données pour avoir une bonne estimation de leur composition en désordre.

La plupart de ces outils calculent des probabilités associées à chaque acide aminé relatives à son inclusion dans une région désordonnée. Les prédictions sont donc binaires, et ne tiennent pas compte de la diversité de types de désordre qu'on peut observer. De plus, ils sont nombreux à considérer l'information évolutive, et seront donc limités pour les séquences ne partageant pas de similitude avec d'autres séquences dans les banques de données.

### *1.2.4.3. AlphaFold2 et HCA, applicables à l'ordre et au désordre ?*

#### **AlphaFold2 et le désordre**

De façon quelque peu inattendue, AF2 s'est également révélé être un bon outil de prédiction du désordre, grâce à la métrique (pLDDT) mise en place pour mesurer le niveau de confiance attribué par résidu. En effet, environ la moitié des résidus présents dans les 11 protéomes initialement ciblés sont associés à de faibles scores pLDDT, et il a été montré que les régions de ce type sont souvent des IDRs (Tunyasuvunakool et al., 2021). Le pouvoir prédictif du pLDDT (ainsi que de la surface accessible au solvant calculée à partir des modèles AF2), a été comparé dans plusieurs études à celui d'autres prédicteurs de désordre (e.g. SPOT-DISORDER2, IUPred2) sur des ensembles de référence, montrant la supériorité de AF2 également dans ce registre (Akdel et al., 2021; Alderson et al., 2022; Binder et al., 2021; Jumper et al., 2021; Tunyasuvunakool et al., 2021; Wilson et al., 2021). Ces différentes études soulignent néanmoins le fait que certaines régions associées à de faibles pLDDT peuvent subir des transitions de l'état désordonné vers un état ordonné (Tunyasuvunakool et al., 2021; Wilson et al., 2021). Wilson et al. (2021) mettent en lumière le fait que des cas potentiels de transition conformationnelle peuvent être suggérés par l'observation de faibles valeurs de pLDDT associées à des éléments de structure secondaire régulière ou de hautes valeurs de pLDDT associées à des régions non structurées. Par ailleurs, Alderson et al., (2022) ont montré qu'une partie significative des IDRs, sièges de transitions conformationnelles, sont prédits par AF2 dans leur conformation repliée, et ceci même si des moules de structures expérimentales de l'état complexé ne sont pas disponibles. Dans tous les cas, il est manifeste qu'AF2 ne peut rendre compte de la plasticité structurale associée aux IDRs et des ensembles conformationnels qui les caractérisent (Alderson et al., 2022).

#### **HCA et le désordre**

Seg-HCA, en identifiant les segments foldables au sein des séquences, peut être considéré comme un prédicteur d'ordre. Les séquences situées en dehors de ces segments foldables sont donc désordonnées, exception faite de certains segments liant des ions. En comparant ces résultats de prédiction de segments foldables avec ceux du prédicteur de désordre IUPred

(Dosztányi et al., 2005b), on s'aperçoit qu'il existe un recouvrement entre les prédictions des deux outils, avec par exemple 13.8% des résidus du protéome humain prédits à la fois en segments foldables (SEG-HCA) et désordonnés (IUPRED) (Figure 1.24 ; Faure & Callebaut, 2013).

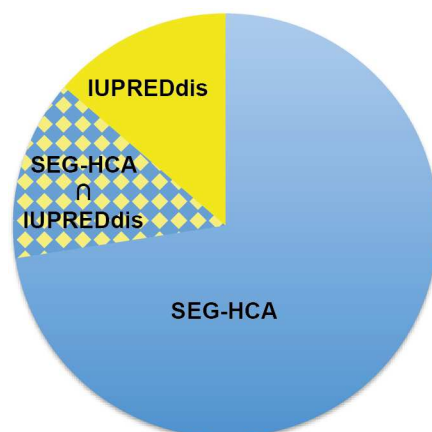


Figure 1.24. Comparaison des prédictions seg-HCA et IUPred. Proportion de résidus prédits comme désordonnés par IUPred long (en jaune), comme foldables par seg-HCA (en bleu) et à la fois comme désordonnés par IUPred et foldables par seg-HCA (mosaïque bleu-jaune). Figure extraite de Faure & Callebaut (2013).

Cet article, ainsi que celui de Bitard-Feildel et al. (2018), montrent qu'une partie importante de ces cas intermédiaires correspondent à des domaines désordonnés à l'état isolé, mais capables d'adopter une structure 3D stable sous contrainte d'un partenaire. Ces domaines peuvent se présenter sous plusieurs formes différentes, selon leur taille et/ou leur composition en amas). Ils peuvent correspondre à des segments foldables courts (des SLiMs ou des MoRFs, cf 1.2.2.2), on en observe un exemple Figure 1.25A. Ces courts motifs conservés sont généralement intégrés dans des régions désordonnées (non foldables, car sans amas hydrophobes), et connaissent des transitions désordre-ordre déclenchées par l'interaction avec des partenaires. Lorsque ces segments désordonnés sont plus longs, on peut observer deux types de tracés HCA différents. Le premier cas représenté Figure 1.25B, montre un tracé HCA moins dense en amas et avec des amas plus petits que ce qu'on observe typiquement dans un domaine globulaire soluble (cas des deux domaines représentés Figure 1.18B). Dans ce cas, le plus faible ratio structure secondaire / boucle explique la difficulté de la séquence à adopter un repliement stable sans l'intervention d'une partenaire extérieur. Enfin, le dernier cas de régions désordonnées capables de se replier sous contrainte est le plus compliqué à discerner des régions ordonnées de repliement stable parce qu'il présente des tracés HCA similaires en termes de densité et de taille des clusters hydrophobes (exemple Figure 1.25C).

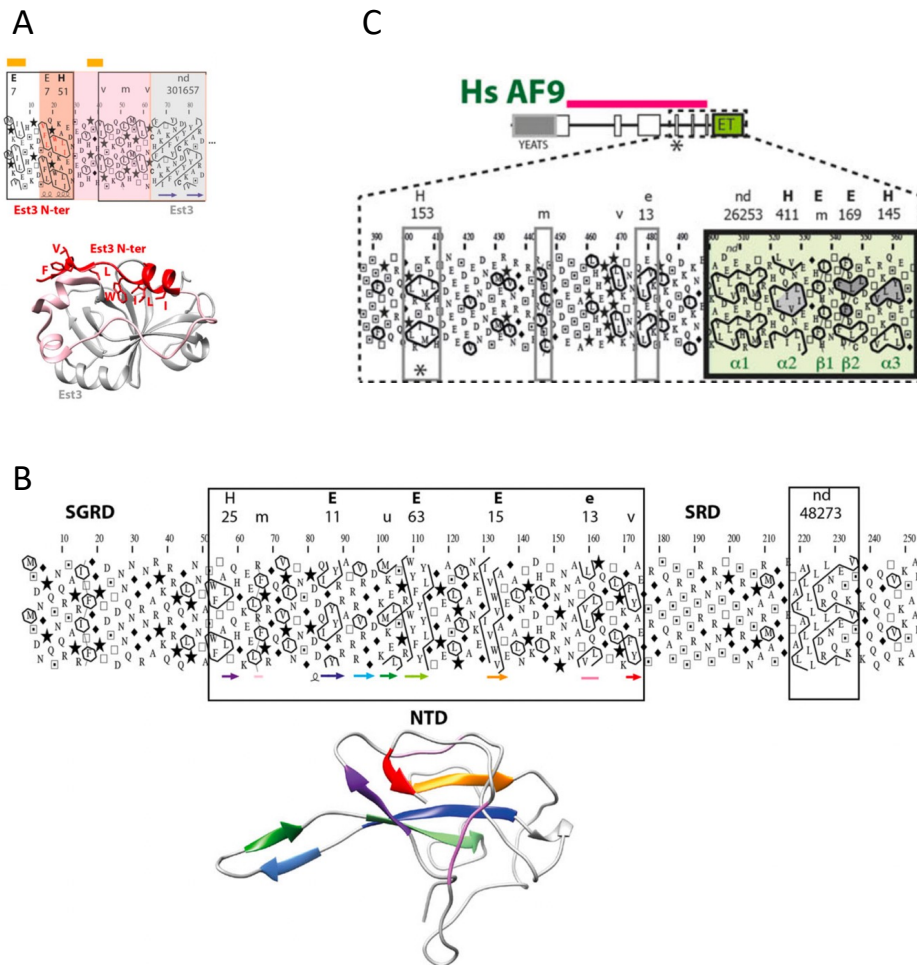


Figure 1.25. Exemples de cas de domaines désordonnés capables de se replier sous contrainte d'un partenaire. A) Cas de segment foldable court dans la protéine *Est3* (UniProt : Q03096, PDB : 2M9V). Le repliement de cette région N-terminale est réalisé grâce à une interaction intra-moléculaire avec le domaine globulaire qui le suit. B) Cas de segment foldable long, peu dense en amas hydrophobes, observé dans la nucléoprotéine du SARS coronavirus humain. C) Cas de segment foldable long, dense en amas hydrophobes dans le domaine ET de la protéine humaine AF9 (UniProt : P42568), qui est désordonné en absence d'interaction avec son partenaire, le peptide AF4 (UniProt : P51825). Figure adaptée de Bitard-Feildel et al. (2018).

En conclusion, les IDRs et IDPs jouent des rôles clés dans les processus cellulaires et sont le siège d'une grande diversité de fonction et d'états structuraux. Il n'existe donc pas un seul type de désordre, mais plusieurs catégories. Il est ainsi difficile de concevoir des prédicteurs prenant en compte à la fois ces caractéristiques globales de désordre et les spécificités propres à chaque catégorie, en particulier sans prendre en compte l'information évolutive. Ceci est d'autant plus vrai que si l'information de désordre, comme celle de l'ordre, est inscrite dans la séquence en acides aminés, l'état de structure observé va également dépendre du contexte dans lequel le segment désordonné se trouve (e.g. présence de charges ou d'amas hydrophobes dans les régions flanquantes) (Bugge et al., 2020).

## 1.3. Les inconnus des protéomes

### 1.3.1. Définitions et quantification

Plusieurs catégories d'inconnu ont été proposées dans la littérature. Malgré l'accroissement important des données de séquençage disponibles, une part significative de la diversité taxonomique au sein du vivant reste inconnue, et les protéomes correspondants restent par conséquent inaccessibles à l'analyse. Certaines régions codantes présentes au sein des génomes ne peuvent être détectées comme telles par les outils conventionnels (e.g. prodigal (Hyatt et al., 2010), NCBI pgap (Li et al., 2021)), en raison de leur taille, de leur biais de composition nucléotidique (Dimonaco et al., 2022), ou d'autres facteurs encore eux-mêmes inconnus. Dans d'autres cas, ce sont les technologies de séquençage qui échouent à la détection ou au traitement de ces régions génétiques (e.g. séquences à fort taux de GC difficiles à séquencer et souvent éliminées ; en cas de duplications de régions génomiques, l'assemblage de ces régions est difficile car l'alignement d'une lecture courte à un seul endroit n'est pas possible) (Ebbert et al., 2019). Dans ces différents cas, on parle d'inconnu inconnu, par définition non quantifiable (Bernard et al., 2018).

Au sein des séquences protéiques prédites à partir des génomes disponibles (que l'on appellera par la suite protéomes), l'inconnu peut concerner des séquences : (i) que l'on ne peut pas rattacher à un taxon connu (sur la base de leur composition nucléotidique, e.g. ANI) (on parle d'inconnu taxonomique) ; (ii) pour lesquelles on ne peut identifier aucune similitude dans les banques de données de (familles de) séquences (on parle de singletons) ; (iii) sans annotation fonctionnelle (on parle d'inconnu fonctionnel) ; (iv) dont la structure 3D n'est pas résolue et/ou non modélisable par modélisation comparative (on parle d'inconnu structural). Ces différentes catégories et leurs quantifications ont été étudiées pour des ensembles de données distinctes (e.g. taxa, environnements) et selon des définitions variables (Almeida et al., 2021; Bernard et al., 2018; Bitard-Feildel et Callebaut, 2017; Perdigão et al., 2015; Thomas et Segata, 2019; Vanni et al., 2022).

Par exemple, la part des données métagénomiques sans annotation fonctionnelle a été quantifiée à 22% d'acides aminés par Vanni et al., (2022), environ 40% de gènes dans les microbiomes humains (Almeida et al., 2021; Thomas et Segata, 2019) et 51.2% de gènes dans l'océan (Carradec et al., 2018). L'inconnu représente également une part non négligeable des données génomiques. En Figure 1.26, on peut en effet observer que 23% des séquences et 47% des résidus d'UniProtKB ne peuvent être annotés fonctionnellement par les profils probabilistes issus de la base de données Pfam (Mistry et al., 2021). Ces valeurs restent stables depuis 2016, malgré la croissance d'UniProtKB.

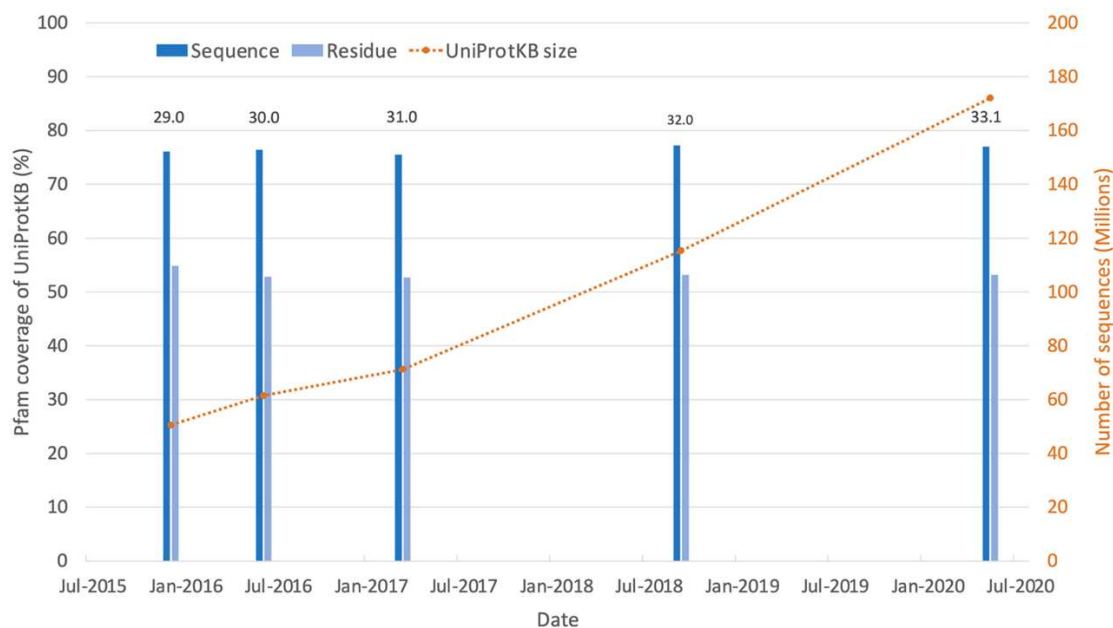


Figure 1.26. Évolution de la couverture d'UniProtKB par Pfam (en séquences et résidus) au cours du temps et de l'enrichissement d'UniProtKB en nouvelles séquences. Figure extraite de Mistry et al. (2021).

Par ailleurs, plus de 20 % des domaines représentés par des profils dans cette banque Pfam correspondent à des domaines de fonction inconnue (DUF pour « domain of unknown function ») (Brown et Babbitt, 2014), alors même que diverses études ont mis en évidence qu'ils étaient présents dans des protéines essentielles et qu'une partie de ceux-ci se retrouvent dans l'ensemble de l'arbre du vivant (Goodacre et al., 2013). Il a été également estimé que plus de 30 % des protéines de *S. cerevisiae* et de l'homme de fonction inconnue correspondent à des enzymes (Ellens et al., 2017). Des travaux d'envergure de génomique structurale ont été entrepris afin de résoudre la structure de membres d'une partie de ces familles DUF (Jaroszewski et al., 2009). Ils montrent que les deux tiers de cas étudiés appartiennent à des branches divergentes de familles de domaines déjà connues et bien caractérisées, alors que d'autres cas semblent correspondre à des nouveaux repliements, même si on y retrouve des fragments présentant des similitudes avec des repliements déjà caractérisés. L'utilisation d'un ensemble de méthodes sensibles de détection de faibles similitudes a par ailleurs conduit à proposer des structures et fonctions potentielles pour presque 20 % des familles DUF, même si ceux-ci présentent des divergences significatives vis-à-vis des motifs fonctionnels, évoquant la présence de vestiges fonctionnels ou l'évolution vers une spécificité différente de substrats ou de partenaires (Mudgal et al., 2015).

L'étude de Perdigo et al. (2015) a quantifié la part d'inconnu structural présent dans la banque de séquences SwissProt (utilisation de l'outil Aquaria (O'Donoghue et al., 2015)), montrant que les régions ou protéines inconnues (nommées dans cette étude respectivement « dark regions » et « dark proteins ») constituent une part variable des protéomes des différents règnes, couvrant par exemple 44 % des acides aminés dans le cas des séquences de protéines d'organismes eucaryotes (Figure 1.27B). L'étude de Bitard-Feildel & Callebaut (2017) a été

réalisée de façon analogue, en analysant les séquences de la banque UniProtKB, et en définissant l'inconnu structural et fonctionnel par la recherche de similitude dans les banques de structures 3D et de familles de domaines (Protein Model Portal, CDD, Pfam). Cette étude montre des résultats similaires à ceux de Perdigão et al. (2015), avec toutefois une part légèrement plus faible de l'inconnu (environ 40 % des acides aminés dans le cas des séquences de protéines eucaryotes) (Figure 1.27C).

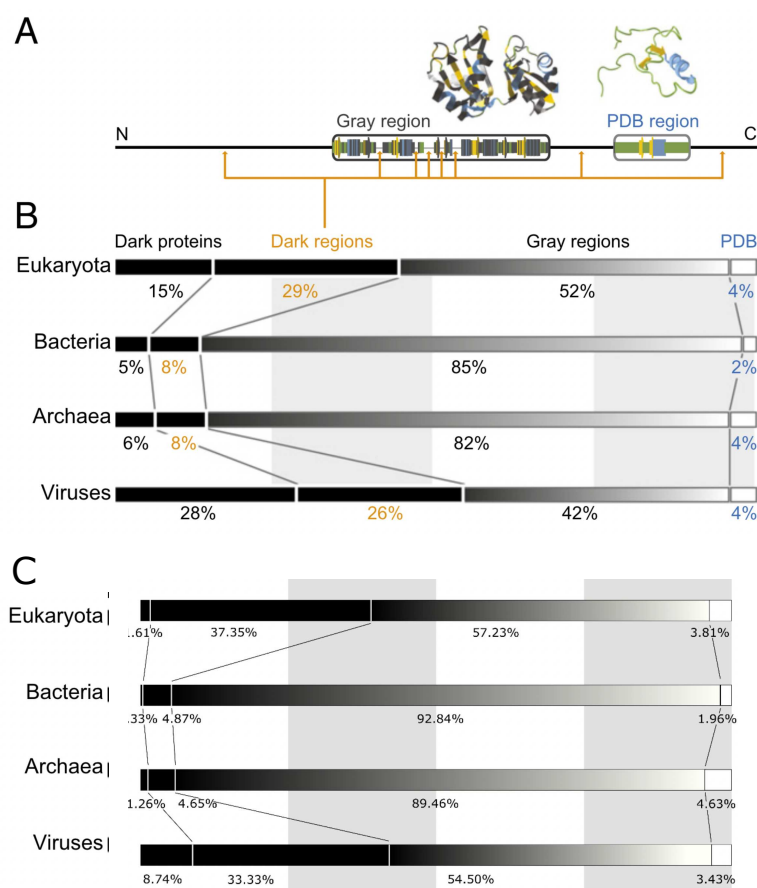


Figure 1.27. Pourcentages d'acides aminés constituant l'inconnu dans les protéines des quatre règnes de la vie, selon deux définitions différentes de l'inconnu. A) Représentation schématique des annotations couvrant une séquence protéique. B) Répartition des résidus dans les protéines de SwissProt en « dark proteins » : protéine sans similitude dans Aquaria ; « dark regions » : régions sans similitude dans Aquaria, « gray regions » : régions présentant des similitudes avec des séquences incluses dans la PDB; « PDB » : séquences de la PDB. C) Répartition de régions dans les protéines d'UniProtKB classées en « dark protein » : protéine non annotée par Pfam, CDD et Protein Model Portal, « dark region » : régions non annotées par Pfam, CDD et Protein Model Portal; « gray region » : régions annotées par Pfam, CDD et Protein Model Portal; « PDB » : séquences de la PDB. Figures A et B extraites de Perdigão et al. (2015). Figure C extraite de Bitard-Feidel & Callebaut (2017).



### 1.3.2. Causes de l'inconnu : erreurs d'annotation, désordre ou nouveauté ?

L'absence d'annotation d'une protéine peut être expliquée de différentes manières. Il se peut qu'il y ait des homologues déjà décrits dans les banques de données, mais que les séquences aient trop divergé pour que des similitudes soient détectables (relations de parenté cachées). Elle peut également résulter de caractéristiques physico-chimiques particulières de la protéine.

Des biais méthodologiques peuvent également survenir au préalable, notamment au moment de prédire la position des séquences codantes (CDS) au sein des séquences génomiques. Les outils de prédiction de CDS font montre de précision quand il est question d'annoter les gènes qui sont véritablement présents (entre 90 et 97% de gènes de procaryotes dont la traduction a été vérifiée expérimentalement sont bien prédits (Hyatt et al., 2010)). Une étude comparative des prédicteurs de CDS sur des génomes procaryotes menée par (Dimonaco et al., 2022) a mis en évidence que les gènes que ces outils échouent à identifier sont de taille significativement réduite par rapport aux gènes bien prédits. En effet, la longueur médiane des gènes non prédits est de 317 nucléotides, contre 837.5 pour celle des gènes bien prédits. Ces biais sur les petits gènes ne s'arrêtent pas aux prédicteurs de CDS, de nombreux outils de bio-informatique ont des limites de tailles dans leurs algorithmes, et ces séquences sont sous-représentées dans les banques de données, alors même que ces petits CDS jouent des rôles décisifs (Andrews et Rothnagel, 2014; Duval et Cossart, 2017; Storz et al., 2014), Le second biais de ces méthodes est la prédiction de CDS là où il ne devrait pas y en avoir, ces prédictions sont appelées des gènes « *spurious* ». Même si leur proportion est très faible, au vu du nombre de séquences recensées dans les banques de données (plus de 100 millions de séquences dans UniProt) cela représente des centaines de milliers de séquences (Höps et al., 2018; Monzon et al., 2022a). Il est donc important de pouvoir identifier ces cas « *spurious* » pour pouvoir limiter leur nombre dans les analyses. Pourtant, les méthodes d'identification des gènes « *spurious* » sont très limitées. La banque de données AntiFam (Eberhardt et al., 2012) recense les séquences qui ont été identifiées manuellement comme « *spurious* ». Celles-ci peuvent être utilisées pour rechercher ces profils de « *spurious* » déjà identifiés dans les jeux de données (méta)génomiques. Enfin, une étude récente a indiqué que si une séquence en acides aminés possède des signatures de repliement, il est peu probable qu'elle soit « *spurious* » (Monzon et al., 2022a); seules 5-20% de séquences aléatoires sont capables d'adopter un repliement tridimensionnel, bien que le contraire ne soit pas vrai.

Différents auteurs assimilent la part d'inconnu des protéomes à des **séquences intrinsèquement désordonnées**, constituant une part importante des protéomes et échappant aux techniques utilisées pour caractériser les structures de domaines repliés (Bhowmick et al., 2016). Les études de Perdigão et al. (2015) et Bitard-Feildel & Callebaut (2017) reportées ci-dessus ont néanmoins démontré qu'une part seulement des séquences non annotées correspondent à du désordre, en utilisant respectivement IUPred et seg-HCA pour prédire les deux états (ordonnés/non ordonnés). Par ailleurs, une partie seulement de l'ordre non annoté mis en évidence dans l'étude de Perdigão et al. (2015) correspondrait à des protéines

membranaires. L'étude de Bitard-Feidel & Callebaut (2017) montre que les segments foldables non annotés ont une propension plus forte vers le désordre et une topologie en résidus hydrophobes différente de ce qui est observée dans le connu (Figure 1.28), indiquant que l'ordre présent dans la part inconnue des protéomes pourrait posséder des caractéristiques sensiblement différentes de celles des domaines repliés annotés.

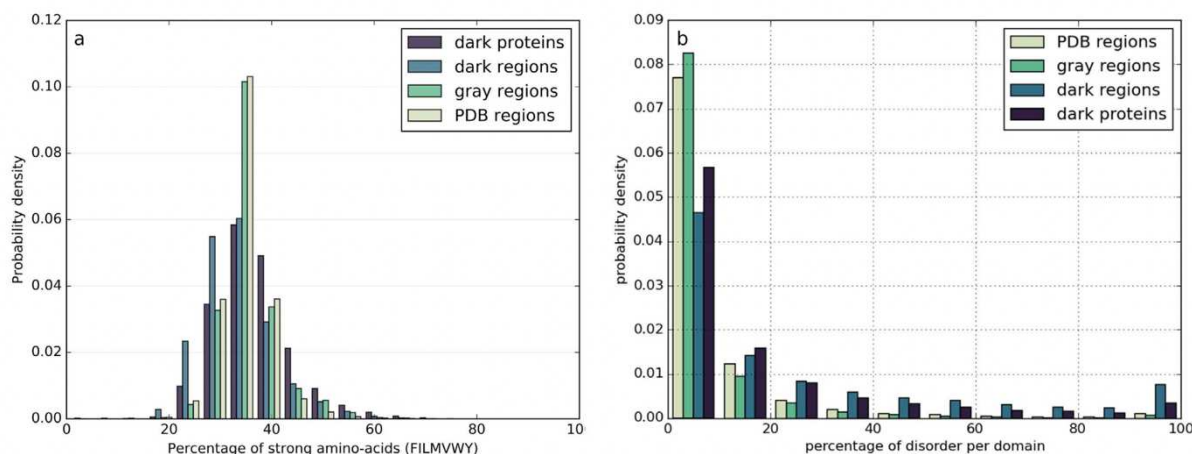


Figure 1.28. Hydrophobicité et désordre. (a) Distributions des fréquences en acides aminés fortement hydrophobes (VILFMYW) des segments foldables dans les différents groupes de séquences protéiques. (b) Distributions de la couverture en désordre dans les segments foldables. Les différentes catégories correspondent à celles décrites en Figure 1.27. Figure extraite de Bitard-Feidel & Callebaut (2017).

Si on écarte la possibilité de parenté éloignée cachée avec des domaines déjà connus (Weisman et al., 2020), on peut émettre l'hypothèse que les séquences ordonnées et sans annotation fonctionnelle correspondent à (i) des séquences d'une famille de protéines qui n'a jamais été décrite d'un point de vue structural et/ou fonctionnel et (ii) des protéines codées par des gènes *de novo* apparus récemment.

L'apparition *de novo* de gènes est un processus d'évolution au cours duquel une portion auparavant non codante du génome devient codante (Oss et Carvunis, 2019). Si le gène traduit est favorable à la survie de l'organisme dans un environnement donné, il sera sélectionné et envahira peu à peu la population où il est apparu. Plus les gènes *de novo* sont récents, plus ils sont restreints à des taxa de bas niveau (Schlötterer, 2015; Tautz et Domazet-Lošo, 2011). Cet événement semble être particulièrement important pour permettre des processus d'adaptation et d'interaction avec l'environnement spécifique à une lignée (Tautz et Domazet-Lošo, 2011). L'identification de gènes *de novo* peut être réalisée selon deux méthodes : (i) l'utilisation de la synténie, permet d'identifier l'apparition d'un gène entre deux gènes dans des génomes proches phylogénétiquement et (ii) l'utilisation de la phylogénie, pour identifier des gènes restreints à une lignée. Les mécanismes conduisant à la naissance et à l'évolution des gènes *de novo* sont sujets à débat. En particulier, deux modèles s'affrontent vis-à-vis des caractéristiques structurales des protéines issues de gènes *de novo* (Papadopoulos et al., 2021). Le modèle de

« préadaptation » stipule que les séquences doivent être suffisamment désordonnées pour ne pas présenter le risque de s'agréger, et être nuisibles à l'organisme (Bitard-Feildel et al., 2015; Ekman et Elofsson, 2010; Schmitz et al., 2018; Wilson et Masel, 2011). Le second modèle indique, au contraire, que les jeunes gènes sont moins enclins au désordre (Carvunis et al., 2012; Vakirlis et al., 2020, 2018).

Vanni et al. (2022) ont proposé une étude de la répartition taxonomique des protéines non annotées, ils montrent notamment que les séquences inconnues sont extrêmement diversifiées, mais phylogénétiquement plus préservées que les séquences annotées fonctionnellement. Les séquences inconnues sont majoritairement restreintes à l'échelle des espèces (Figure 1.29). Une hypothèse découlant de ces observations est que l'inconnu est enrichi en gènes accessoires. Une étude de Méheust et al. (2022) corrobore cette hypothèse, en démontrant l'importance des gènes accessoires dans l'inconnu identifié au sein de génomes d'archaea.

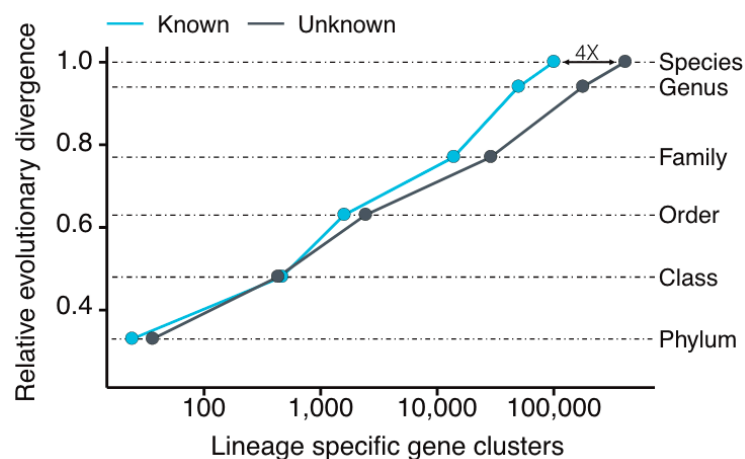


Figure 1.29. Distributions des clusters de gènes d'Agnostos-DB (cf 1.3.3) spécifiques à une lignée en fonction de l'échelle taxonomique. Figure extraite de Vanni et al. (2022).

### 1.3.3. Ressources et outils.

Les premières séquences inconnues recensées comme telles dans une banque de données datent de 2001 (Galperin et Koonin, 2004), avec la création du label *hypothetical proteins*, qui désigne les protéines pour lesquelles aucune similitude n'a été identifiée avec les séquences de protéines déjà décrites. Les banques de données fonctionnelles telles que Pfam et InterPro ont par la suite intégrées l'étiquette « DUF », mentionnée précédemment, pour désigner les domaines dont la fonction est inconnue.

Avec l'essor récent de la métagénomique, et les développements d'outils permettant l'annotation séquences en masse, des banques de données proposant des classifications de l'inconnu commencent à émerger.

C'est le cas d'Agnostos-DB (Vanni et al., 2021), qui renferme des séquences de génomes de référence issus de GTDB, banque de données de génomes procaryotes qui propose une

classification taxonomique différente de celle du NCBI (Parks et al., 2022) et de données métagénomiques océaniques (TARA, Malaspina (Acinas et al., 2021), OSD (Kopf et al., 2015), GOS (Yooseph et al., 2007)) et des microbiomes humains (HMP (Parks et al., 2018)). Agnostos-DB recense 6 572 081 protéines, classées en : (K) *Known* pour les séquences de fonction annotée par Pfam, (KWP) *Known without Pfam* pour les séquences annotées fonctionnellement par une autre banque de données, (GU) *Genomic Unknown* pour les séquences de fonctions inconnues mais retrouvées dans les données génomiques et enfin, (EU) *Environmental Unknown* pour les séquences métagénomiques qui ne trouvent pas d'homologie dans les banques de données génomiques (Figure 1.30). La classification est réalisée par l'algorithme Agnostos (Vanni et al., 2022), qui combine des outils de classification non supervisée et de recherche d'homologie lointaine pour l'annotation fonctionnelle.

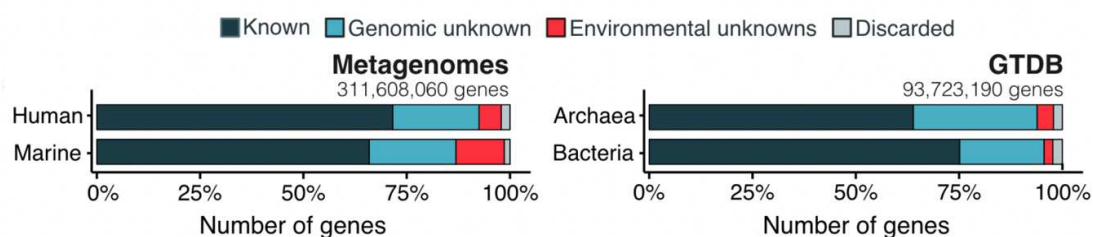


Figure 1.30. Quantification de chaque classe d'Agnostos-DB pour les données métagénomiques (à gauche) et génomiques (à droite). Figure extraite de Vanni et al. (2022).

La banque Novel Gene Families (Río et al., 2022) regroupe 413 335 familles de protéines de fonction inconnue et/ou de taxa inconnu. Les séquences sont issues de catalogues multi-habitats, de métagénomiques du microbiome humain, d'échantillons océaniques et de données génomiques (GTDB). De manière similaire à Agnostos, elle combine classification non supervisée et recherche d'homologie lointaine pour identifier ces familles.

Enfin, Wyman et al. (2018) proposent une liste de 6 668 protéines non annotées et conservées de microorganismes (appelées FUnkFams). Ces séquences sont issues de TARA Ocean et de HMP. Leur identification repose sur une classification non supervisée, mais cet outil a l'inconvénient de ne pas procéder à une recherche d'homologie lointaine, le nombre de familles considérées comme inconnues risque donc d'être largement surestimé. Comme l'a montré Godzik (Godzik, 2011), l'utilisation d'algorithmes de recherche d'homologie lointaine peut permettre de réduire de 70% le nombre de familles protéiques pour lesquelles on ne trouve pas de similitude dans les banques de données fonctionnelles.

Ces banques de données présentent l'avantage de proposer de très grands jeux de données qui ont déjà passé les étapes d'annotation en fonction et en homologues, très coûteuses en temps de calcul. Cependant, il est difficile de mener des recherches exploratoires sur les protéines de l'inconnu, car il est impossible de savoir précisément quelles régions protéiques sont couvertes par des annotations fonctionnelles ou par de l'homologie, et lesquelles ne le sont pas.

Le prédicteur de structures protéiques AlphaFold2 (AF2) permet d’aller plus loin dans l’annotation des séquences, comme le montrent Porta-Pardo et al. (2022) qui indiquent que l’application d’AF2 au protéome humain a permis de réduire la part d’inconnu (hors séquences prédites comme désordonnées) de 26% à 10% (Figure 1.31). L’apport d’AF2 est renforcé par la modélisation de toutes les protéines d’UniProtKB (plus de 200 millions). Cependant, seules 35% auraient une prédiction confiance élevée et 45% une confiance suffisante pour beaucoup d’applications (Callaway, 2022).

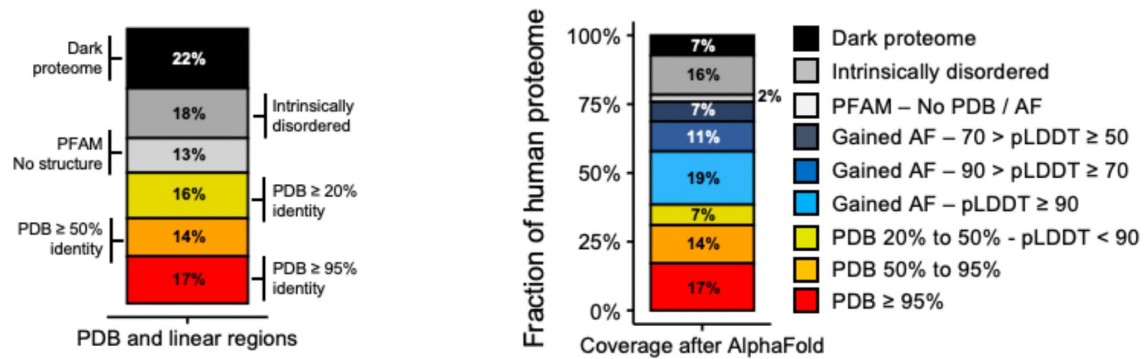


Figure 1.31. Couverture du protéome humain en structures 3D et en domaines Pfam avant (à gauche) et après (à droite) l’apport d’AF2. Figure adaptée à partir de Porta-Pardo et al. (2022).

Les structures prédites avec une confiance très faible correspondent aux séquences de l’inconnu structural. Si ces prédictions renferment une part importante de désordre (cf 1.2.4.3), il est également possible qu’elles contiennent de l’ordre qu’AF2 n’a pas su prédire par manque d’information évolutive et/ou en raison de motifs structuraux encore non observés. Or, on les observe en quantité non négligeable dans le protéome de l’homme (28%), et en quantité réduite mais tout de même significative dans le protéome bien étudié de l’espèce bactérienne *Escherichia coli* (2.9%) (Thornton et al., 2021).

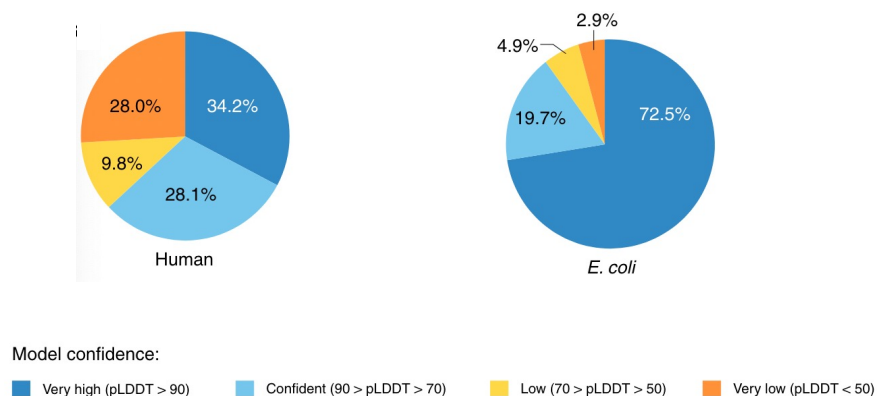


Figure 1.32. Catégorie pLDDT par résidus de la prédiction structurale par AlphaFold2 des protéomes d’Homme et d’*Escherichia coli*. Figure adaptée de Thornton et al., (2021).

Ainsi, même si AF2 permet de progresser dans l’annotation structurale et fonctionnelle, une part d’inconnu subsiste (prédictions de confiance très faible), correspondant à des séquences a priori non désordonnées.

## 1.4. Stratégie et objectifs de la thèse :

Comme nous l'avons vu dans cette introduction, l'annotation fonctionnelle des séquences de protéines reste un défi, malgré la quantité croissante de données disponibles à partir des (méta)génomés. En effet, une fraction significative des protéomes n'est toujours pas annotée, laissant encore inaccessible une partie du répertoire fonctionnel du vivant, y compris des innovations moléculaires ayant une valeur thérapeutique ou environnementale. Le manque d'annotation fonctionnelle est en partie dû aux limites des approches actuelles dans la détection de relations cachées au sein des séquences, ou à des caractéristiques spécifiques telles que le désordre. En outre, la prédiction de la fonction des gènes orphelins ou ayant récemment émergé à partir de régions génomiques préalablement non codantes (appelés gènes *de novo*) est un défi fondamental.

Dans ce contexte, l'objectif de ma thèse était de développer des approches méthodologiques basées sur les signatures structurales des domaines protéiques repliés, afin de caractériser davantage les séquences protéiques dont la fonction est actuellement inconnue (également appelé le « dark proteome »), y compris en l'absence d'informations évolutives.

J'ai tout d'abord développé un système de score, désormais implémenté dans l'outil pyHCA développé au laboratoire, afin d'estimer la capacité d'une séquence d'acides aminés à se replier sous la forme d'une structure 3D. Cette métrique, ou score HCA, est basée sur la densité de séquence en amas hydrophobes (HC), qui correspondent principalement à des structures secondaires régulières. J'ai optimisé ce système de score à partir de l'analyse de bases de données de référence de l'ordre (SCOPE, OPM) et du désordre (DisProt). En appliquant ce score aux séquences de segments repliables (i.e. correspondant aux îlots denses en HC), j'ai pu décrire le continuum entre ordre et désordre, couvrant divers états allant des conformations étendues (« random coils ») aux globules fondus et caractériser des cas d'ordre conditionnel. Ces travaux sont décrits dans le Chapitre 2.

Dans un second temps, j'ai combiné ce score HCA avec les prédictions de structure 3D d'AlphaFold2 (AF2) disponibles pour 21 protéomes de référence, allant des procaryotes aux eucaryotes unicellulaires et multicellulaires avec différents modes de vie. Une grande partie des acides aminés modélisés par AF2 avec un très faible score de confiance sont inclus dans des segments non repliables, ce qui confirme la qualité d'AF2 en tant que prédicteur du désordre. Cependant, au sein de chaque protéome, de longs segments (> 30 acides aminés) avec un score de confiance AF2 très faible présentent également des caractéristiques de domaines solubles repliés, telles qu'estimées par pyHCA. Cela suggère un ordre caché (conditionnel ou inconditionnel), qui n'est pas détecté par AF2 en raison du manque d'information évolutive, ou correspondant à des modèles de repliement non répertoriés. Ces travaux sont décrits dans le Chapitre 3.

Enfin, j'ai réalisé une exploration préliminaire des protéines ou régions non annotées en utilisant pyHCA, présentée en Discussion (Chapitre 4). J'ai collecté des jeux de données grâce au développement et à l'application d'une nouvelle procédure d'annotation. Bien que ces séquences non annotées soient enrichies en désordre, une partie importante d'entre elles présente des caractéristiques de type globulaire soluble. Ces séquences seraient de bons candidats à valider et à caractériser expérimentalement. De plus, l'analyse de gènes *de novo* validés expérimentalement m'a permis de contribuer au débat actuel sur les caractéristiques structurales des protéines codées par ces gènes, en mettant en évidence un enrichissement en désordre et une grande diversité d'états structuraux.

Le dernier chapitre de ce manuscrit de thèse présente les perspectives offertes par ce travail, en particulier pour l'identification de cibles porteuses d'innovations structurales et/ou fonctionnelles. Ces perspectives sont replacées dans le contexte d'une recherche foisonnante dans le domaine de la prédiction des structures et fonctions des protéines et de leurs origines, prenant en particulier appui sur les possibilités offertes par les méthodes d'apprentissage pour exploiter les très grandes masses de données actuellement disponibles.





**Chapitre 2. DEVELOPPEMENT DU  
SCORE HCA ET COMBINAISON AVEC LES  
PREDICTIONS D'ALPHA FOLD2 POUR  
L'ETUDE DU CONTINUUM ORDRE/DESORDRE**

## Présentation de l'article

L'article présenté dans ce chapitre a été publié dans *PROTEINS*.

Comme vu dans l'introduction, il existe plusieurs méthodes pour prédire le contenu en ordre/désordre des séquences protéiques (e.g. (Dass et al., 2020; Dosztányi et al., 2005b; Galzitskaya et al., 2006; Hanson et al., 2019; Hu et al., 2021; Liu et al., 2019; Mirabello et Wallner, 2019; Necci et al., 2021; Romero et al., 1997; Wang et al., 2016; Tuo Zhang et al., 2012)). Le prédicteur de structures 3D AlphaFold2 (Jumper et al., 2021) a introduit le score pLDDT pour évaluer la confiance accordée à ses prédictions, mais il a également été montré qu'il pouvait être utilisé comme outil de prédiction du désordre au sein des protéines, cet état étant globalement associé à de faibles valeurs de score pLDDT (Akdel et al., 2021; Ruff et Pappu, 2021; Wilson et Masel, 2011). Cependant, ces prédicteurs proposent généralement une classification binaire (ordre ou désordre), et ne permettent pas d'apprécier la nature du désordre (e.g. séquence complètement désordonnée, séquence fluctuant entre ordre et désordre, séquence capable de se replier partiellement ou totalement sous contrainte, ...) (van der Lee et al., 2014). De plus, la plupart de ces méthodes reposent sur l'information évolutive, rendant impossible la prédiction de séquences de l'inconnu génomique, pour lesquelles on ne retrouve pas d'homologues dans les banques de données.

Je présente dans ce chapitre le score HCA, une métrique reflétant le potentiel de repliement d'une protéine (ou « foldabilité »). Ce score repose uniquement sur l'information de séquence, et plus précisément sur sa densité en acides aminés hydrophobes et en amas hydrophobes, tels qu'ils sont définis par l'approche HCA (cf 1.1.4.4). Pour tenir compte de l'hétérogénéité des protéines, le score n'est pas calculé sur sa séquence entière mais sur ses segments foldables, définis par la fonction *segment* du module pyHCA. Cela nous permet de préciser les différentes textures présentes au sein de la protéine et de la segmenter en régions structurellement homogènes. Le but de ce travail est de proposer un outil permettant de positionner les segments foldables d'une protéine sur le continuum désordre/ordre. Aux extrémités de ce continuum on retrouve, d'une part, des séquences repliées très compactes (présentant des boucles de petite taille et peu nombreuses), et d'autre part, des protéines complètement désordonnées, de structure étendue (segments non foldables).

Pour que le score HCA permette de discriminer au mieux le désordre de l'ordre, les paramètres qui entrent en jeu dans son calcul sont optimisés par validation croisée k-fold sur les séquences de trois banques de données de référence recensant des séquences dont le statut d'ordre et de désordre a été déterminé expérimentalement. Les banques de données utilisées sont DisProt v8.0.2 (Hatos et al., 2020) pour les régions désordonnées, SCOPe v2.0.7 (Fox et al., 2014) pour les domaines globulaires solubles et OPM (Lomize et al., 2006) pour les domaines transmembranaires. L'optimisation a été réalisée sur les segments foldables de ces séquences, qui couvrent 90% et 96% de SCOPe et OPM, respectivement, et uniquement 50% de DisProt.

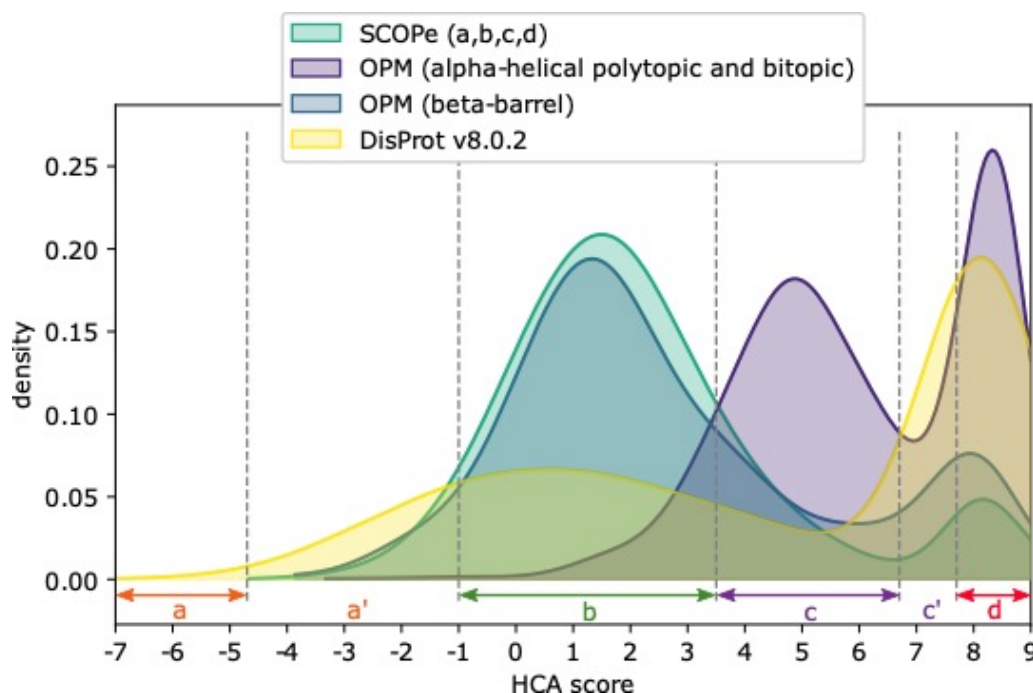


Figure 2.1. Distribution des scores HCA des segments foldables définis sur les séquences des banques de données SCOPe, OPM (hélices alpha bitopiques et polytopiques et tonneaux beta) et DisProt, à partir de laquelle on définit quatre types de segments foldables. *a* : désordre, *b* : globulaire soluble, *c* : membranaire (hélices alpha), *d* : segments à un amas hydrophobe

Le score HCA obtenu après optimisation permet d'évaluer globalement le rapport ordre/désordre dans une séquence protéique, et, en parcourant la gamme de scores HCA, nous observons en effet une continuité d'états structuraux. Nous délimitons ainsi quatre zones (notées *a,b,c,d*) à partir de la distribution de scores HCA (Figure 2.1). La zone (*a*) est caractérisée par des scores inférieurs à -4.7, et inclut des segments qui font preuve d'une capacité à fluctuer entre désordre et ordre et/ou à interagir avec des partenaires. La zone (*b*) est délimitée par l'intervalle [-1,3.5], nous y retrouvons très majoritairement les segments foldables de domaines globulaires solubles. Nous y retrouverons également la plupart des segments foldables membranaires de type tonneaux beta et des cas d'ordre conditionnel. Ces cas de désordre correspondent à des segments foldables non repliés ou partiellement repliés et capables d'une transition désordre-vers-ordre après liaison à un partenaire, ou encore de segments repliés mais instables en l'absence d'éléments qui les stabilisent. La zone (*c*), entre [3.5,6.7], abrite très majoritairement les segments foldables de domaines membranaires de type alpha. Les scores supérieurs à 7.6 constituent la zone (*d*), ils caractérisent des segments foldables souvent courts, couverts par un seul amas hydrophobe. Ils correspondent à des hélices alpha de domaines membranaires bitopiques (un seul passage dans la membrane), ou à des motifs de reconnaissance associés à l'état libre à du désordre (MoRFs et SLiMs).

Nous observons donc des segments désordonnés tout au long de la gamme de scores HCA (de -7 à 9), ce qui reflète bien leur diversité, et permet de distinguer les différents types de désordre. Ce comportement contraste avec celui des domaines globulaires solubles ou membranaires, pour lesquels les gammes de valeurs de scores HCA sont plus étroites.

Nous avons appliqué ensuite le découpage en segments foldables et le calcul du score HCA aux prédictions d'AlphaFold2 (AF2) réalisés pour 21 protéomes de référence, recensés dans *AlphaFold protein structure database* (AFDB) v1. Les prédictions d'AF2 sont classées en quatre catégories selon leur niveau de confiance, déterminée par le score par résidu pLDDT. Pour un  $pLDDT < 50$  les prédictions sont considérées de très faible confiance, de faible confiance pour  $50 < pLDDT < 70$ , de confiance élevée  $70 < pLDDT < 90$ , et de confiance très élevée si  $pLDDT > 90$ .

Nous observons que l'analyse HCA (découpage en segments et calcul du score HCA) et le pLDDT d'AF2 sont globalement en concordance. En particulier, les prédictions d'AF2 de faible et très faible confiance sont majoritaires dans les segments non foldables (couverts à 61.3% de résidus de faible confiance et 17.1% de très faible confiance) (Figure 4 de l'article), et dans les segments foldables de scores HCA typiques du désordre ( $[-7, -4]$  et  $[7, 9]$ ) (Figure 5). Au contraire, les segments foldables de scores HCA typiques des domaines globulaires solubles ou membranaires ( $[-1 ; 6]$ ) sont enrichis en prédictions de haute et très haute confiance (représentant respectivement 45.7% et 27.1% des résidus) (Figure 5).

Cependant, nous observons quelques désaccords apparents entre les prédictions d'AF2 et HCA (Figure 8). Dans le non foldable et le foldable typique du désordre, nous observons des segments prédits avec une confiance très élevée, correspondant à des petits segments isolés ou à des cas particuliers où le repliement n'est pas induit par un cœur hydrophobe, et qui échappent donc au découpage de pyHCA, mais pas à la prédiction AF2. Dans le foldable correspondant aux domaines globulaires solubles, nous retrouvons des prédictions de très mauvaise qualité. Un exemple est celui de MbiA (Figure 7.a), une protéine d'*E. coli*, qui posséderait des structures secondaires régulières, comme suggéré par la présence de plusieurs amas hydrophobes sur le tracé HCA, mais qu'AF2 n'est pas capable de prédire. Ces derniers résultats indiquent qu'une prédiction de faible pLDDT ne correspond pas forcément à du désordre.

Cette étude a été complétée par une analyse du lien entre confiance de prédiction et score HCA à l'échelle des différents protéomes, mettant en évidence les variations et les spécificités taxonomiques (Figure 6). Nous observons ainsi que les protéomes d'eucaryotes contiennent moins de segments foldables et que les prédictions 3D des segments foldables de scores HCA typiques de domaines globulaires solubles présentent des scores pLDDT plus faibles que les protéomes de procaryotes, à l'exception de *Plasmodium falciparum*. Cet eucaryote unicellulaire présente le même taux de segments foldables que les autres (80-85%), mais la couverture en prédictions de très faible confiance par AF2 est plus importante (39.7% des résidus en segments foldables de *P. falciparum* ont de score pLDDT très faible, alors que cette valeur varie entre 14.4% et 23.5% pour les autres eucaryotes).

En conclusion, nous proposons ici un outil permettant de caractériser le potentiel de repliement et les propriétés structurales de segments protéiques, sans utiliser l'information évolutive. Cet outil peut donc être appliqué à des séquences sans homologues présents dans les

banques de données de séquences, et contribuer à réduire la part d'inconnu génomique (voir Chapitre 4).

Pour distinguer les segments foldables désordonnés qui ont des scores similaires à ceux observés pour des domaines globulaires solubles, d'autres caractéristiques pourraient être considérées. Nous avons commencé à explorer quelques pistes. Il serait intéressant de considérer (i) la taille des segments foldables, ceux présents dans le désordre étant souvent plus petits que ceux présents dans des séquences ordonnées (seules 48.4% des séquences de DisProt sont de taille supérieure à 30 résidus, contre 85.1% des segments de SCOPe); (ii) la composition en acides aminés des séquences, les segments désordonnés étant enrichis en résidus polaires; (iii) les caractéristiques (taille, propriété des acides aminées) des amas hydrophobes des segments foldables, et des régions liant les amas hydrophobes entre eux.

## **A sequence-based foldability score combined with AlphaFold2 predictions to disentangle the protein order/disorder continuum**

Apolline Bruley, Tristan Bitard-Feidel, Isabelle Callebaut\*, Elodie Duprat\*

Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, 75005 Paris, France

\* Isabelle Callebaut and Elodie Duprat should be considered joint senior authors

([isabelle.callebaut@sorbonne-universite.fr](mailto:isabelle.callebaut@sorbonne-universite.fr), [elodie.duprat@sorbonne-universite.fr](mailto:elodie.duprat@sorbonne-universite.fr))

## ABSTRACT

Order and disorder govern protein functions, but there is a great diversity in disorder, from regions that are – and stay – fully disordered to conditional order. This diversity is still difficult to decipher even though it is encoded in the amino acid sequences. Here, we developed an analytic Python package, named *pyHCA*, to estimate the foldability of a protein segment from the only information of its amino acid sequence and based on a measure of its density in regular secondary structures associated with hydrophobic clusters, as defined by the Hydrophobic Cluster Analysis (HCA) approach. The tool was designed by optimizing the separation between foldable segments from databases of disorder (DisProt) and order (SCOPE (soluble domains) and OPM (transmembrane domains)). It allows to specify the ratio between order, embodied by regular secondary structures (either participating in the hydrophobic core of well-folded 3D structures or conditionally formed in intrinsically disordered regions) and disorder. We illustrated the relevance of *pyHCA* with several examples and applied it to the sequences of the proteomes of 21 species ranging from prokaryotes and archaea to unicellular and multicellular eukaryotes, for which structure models are provided in the AlphaFold2 databases. Cases of low-confidence scores related to disorder were distinguished from those of sequences that we identified as foldable but are still excluded from accurate modeling by AlphaFold2 due to a lack of sequence homologs or to compositional biases. Overall, our approach is complementary to AlphaFold2, providing guides to map structural innovations through evolutionary processes, at proteome and gene scales.

## KEYWORDS

Protein foldable segments, Hydrophobic Cluster Analysis, soluble and transmembrane domains, IDPs/IDRs, AlphaFold protein structure database

## ABBREVIATIONS

aa : amino acids, AF2 : AlphaFold2, AFDB : AlphaFold Protein Structure DataBase, CAID : Critical Assessment of protein Intrinsic Disorder, DisProt : dabase of intrinsically Disordered Proteins, HCA : Hydrophobic Cluster Analysis, IDPs/IDRs : Intrinsically Disordered Proteins/Regions, OPM : Orientations of Proteins in Membrane, PDB : Protein Data Bank, pLDDT : predicted Local Distance

Difference Test, RSSs : Regular Secondary Structures, SCOPe : Structural Classification of Proteins – extended, TM : Transmembrane

## INTRODUCTION

Protein order has been largely explored by experimental approaches so that the protein fold universe has been widely mapped<sup>1-3</sup>. This has led to a comprehensive inventory of the combinations according to which regular secondary structures (RSSs) are assembled to form compact, well-organized 3D structures, associated with specific functions<sup>4</sup>. Over the years, the structure-function paradigm has however evolved to integrate intrinsically disordered proteins/regions (IDPs/IDRs), which lack well-defined 3D structures under physiological conditions but fulfill a variety of functions, in particular in signaling and regulatory pathways<sup>5-8</sup>. IDPs/IDRs are characterized by a heterogeneous spatiotemporal structural organization<sup>9</sup>. They correspond to very diverse entities, from highly extended, heterogeneous unstructured states to compact but disordered molten globules<sup>10</sup>. They include short linear motifs and longer regions promoting molecular recognition and protein-protein interactions, which have led to elaborate specific classification schemes related in particular to their amino acid sequence characteristics<sup>6,11</sup>. IDPs/IDRs were also shown to play a key role in the formation of higher-order assemblies and in the control of many cellular processes via their participation in biomolecular condensates, through multivalent interactions leading to liquid-liquid phase separation<sup>12,13</sup>. IDPs/IDRs are generally defined by a heterogeneous ensemble of conformations, undergoing rapid interconversion<sup>14,15</sup>; some can fold into a unique conformation upon binding with a partner or within oligomeric complexes<sup>16</sup>, while some cases of disorder maintained in the bound state were also described<sup>17</sup>.

Characterization of IDPs/IDRs is challenging as they are generally “unseen” by traditional structural biology methods and are therefore considered as the dark side of protein universe<sup>18</sup>. Their identification at large scale relies on computational methods that predict them directly from the information of the amino acid sequence<sup>19</sup>. Some predictors are trained on experimental annotations of protein disorder, as stored in the DisProt database<sup>20</sup>, while others are not, relying on physicochemical properties and predicting disorder as lack of, or deviation from order<sup>21,22</sup>. Using such predictive tools, IDPs/IDRs were found abundant at the proteome level, making up approximately 30 % of residues in the human proteome and up to 50 % in some unicellular eukaryotes such as parasitic protozoa, enriched in long IDRs (with at least 30 consecutive disordered residues)<sup>23-26</sup>. In contrast, the fractions of disordered residues represented less than 28% in most archaeal and bacterial proteomes. The quality of disorder predictors has improved over time, in particular by considering deep learning techniques and evolutionary information, as



illustrated in the recent Critical Assessment of protein Intrinsic Disorder (CAID)<sup>27</sup>. The predicted Local Distance Difference Test (pLDDT) introduced in the recent AlphaFold2 (AF2) predictor, a deep learning program which predicts 3D structures with an unprecedented accuracy<sup>28</sup> and which was applied at proteome scale<sup>29</sup>, was also shown to provide a good metric for identifying order and disorder<sup>30-32</sup>.

Only a few computational approaches have addressed the issue of the predictions of different, multiple states (or flavors) of IDPs/IDRs, while not considering evolutionary information (*e.g.*<sup>33,34</sup>). Here, we propose an approach for appreciating the disorder/order degree in a protein segment from the only information of its amino acid sequence, which is based on a measure of the overall density in regular secondary structures, as predicted through Hydrophobic Cluster Analysis (HCA). HCA-based hydrophobic clusters match the positions of regular secondary structures constituting the building blocks of folded domains<sup>35-39</sup>. The hydrophobic alphabet and rules used for the definition of hydrophobic clusters have been supported by comparison with experimental data, and the method has been successfully applied, for instance, to the identification of remote relationships based on the conservation of 2D signatures associated with hydrophobic clusters (see **Supplementary Dataset S1** for a complete description of the method). The use of a simple hydrophobic/non-hydrophobic dichotomy (rather than an hydrophobicity scale), associated with the use of a two-dimensional net for defining the amino acid neighborhood, offers an efficient way to reveal these signatures in remotely related sequences<sup>40</sup>. A tool, called *SEG-HCA*, was previously developed for the delineation of regions with a high density in hydrophobic clusters, which have been shown to correspond to domains which have the ability to fold, either in an autonomous way or upon contact with partners<sup>41,42</sup>. Contrasting with otherwise performant methods based on evolutionary couplings (*e.g.*<sup>43</sup>) or even earlier tools based on propensities to be in ordered or disorder states (*e.g.*<sup>44</sup>), the advantage of *SEG-HCA* is to allow the prediction of foldable domains from the only information of a single amino acid sequence, without the prior knowledge of homologous sequences or consideration of pre-calculated propensities. Hence, by identifying structurally homogeneous entities (*i.e.* the foldable segments), this approach allows to highlight, in absence of evolutionary information, elements that are likely to interact, however without being able to predict how they interact (which is possible by taking into account residues co-evolution). At proteome scale, *SEG-HCA* “order” predictions totalize more amino acids than the “undisordered” predictions performed by the popular IUPRED tool<sup>45,46</sup>, which captures the inter-residue interaction capacity by energy estimation<sup>42</sup>. The overlap between order and disorder detected by this comparison concentrates small sequences that are able to undergo disorder-to-order transitions<sup>42</sup>.

We optimized the residue weights of the here-proposed density metric for an optimal separation of foldable domains found in reference databases of order (soluble and transmembrane domains extracted from SCOPe and OPM, respectively) and disorder (DisProt.), thereby deconvolving the spectrum of disorder according to the order/disorder ratio and specifying different types of disorder. Using this scoring scheme, we analyzed the per-residue confidence (pLDDT) scores of AF2 structural predictions in 21 reference proteomes, for species ranging from prokaryotes and archaea to unicellular and multicellular eukaryotes with different lifestyles<sup>29,47</sup>. Our analysis provides new elements to distinguish cases where low-confidence structural predictions are indeed related to disorder, as now commonly reported<sup>21,22</sup>, from those of domains which are foldable but whose structures cannot be accurately predicted due to AF2 intrinsic limitations. Overall, the complementarity of *pyHCA* and AF2 provides guides to map structural innovations through evolutionary processes, at proteome and gene scales.

## MATERIALS AND METHODS

### Delimitation of foldable segments within protein sequences

The HCA methodology is described in details in **Supplementary Dataset S1**. *SEG-HCA*<sup>42</sup> was previously developed to automatically delineate regions with high density in hydrophobic clusters, constituting potential “foldable” domains within protein sequences. The new version of *SEG-HCA* was rewritten for speed and includes new functionalities, including the calculation of HCA score (see below). In addition, we rewrote and packaged in a python module the tools *SEG-HCA* (in a function named *segment*), *TREMOLLO-HCA* (Traveling through REMOte homoLOgy)<sup>48</sup> and the HCA drawing, as well as provided some utility scripts to analyze their results, adding information on amino acids conservation and taxonomy. These tools can significantly help detection of hidden relationships between sequences, as repeatedly performed using the HCA approach in an expert-based way (see references in<sup>41</sup> and **Supplementary Dataset S1**). The package can be used as a Python library, named *pyHCA*, or as a standalone tool, named *HCAtk*, and is provided at <https://github.com/DarkVador-HCA/pyHCA> under the CeCILL-C license agreement.

### HCA score implementation

The HCA score was introduced in order to describe the general composition in hydrophobic clusters and hydrophobic amino acids in foldable segments. Each residue of a protein sequence is associated with a class regarding the residue type and hydrophobicity. Such a residue is either (i) in a hydrophobic cluster

and hydrophobic according to the common HCA alphabet, which considers as strong hydrophobic the seven amino acids V, I, L, M, F, Y, and W, (ii) in a hydrophobic cluster and non-hydrophobic, or (iii) outside a hydrophobic cluster. A value is attributed to each class and the HCA score ( $S_{HCA}$ ) is computed as follow:

$$S_{HCA} = \sum_{i=1}^N w_{HCA}(seq_i, class)/N. \quad (1)$$

with  $w_{HCA}(seq_i, class)$  the weight associated with the  $i$  amino acid of a sequence from a given class, and  $N$  the sequence length.

Therefore, the HCA score scales with the density in hydrophobic clusters and in strong hydrophobic residues inside the clusters. As the HCA score calculation motivation is to provide an estimation of the globular character, *i.e.* the foldability of a domain, the weight of each residue was optimized inside each of the three classes to minimize the overlap between the distributions of the HCA scores computed on non-redundant disordered foldable segments from DisProt v8.0.2 sequences and non-redundant foldable segments of globular proteins from the SCOPe (soluble domains) and OPM databases (transmembrane domains), respectively (see below for details on the datasets). During the optimization step, the allowed possible values of the weights were discrete inside the  $[-10; 10]$  interval ( $[-1; 1]$  scaled to one order of magnitude for better readability). For a given combination of weights, the distribution overlap (*i.e.* the criteria to be minimized) was estimated by histogram intersection as follows:

$$\sum_{i=1}^B \min(H_1(i), H_2(i)),$$

with  $B$  the number of histogram bins (set to 60), and  $H_1(i)$  and  $H_2(i)$  the values in normalized histograms of HCA score for non-redundant dataset Disprot v8.0.2 and non-redundant dataset SCOPe and OPM for bin  $i$ , respectively. The optimization was achieved using 10-fold cross-validation, repeated 10 times. We selected the set of parameters that was the most frequently the optimal one out of the 10 iterations, and that allowed to achieve the lowest overlap between disorder and order.

This implementation is an updated version (v2) of a previous one, described in <sup>49</sup> and applied in investigations such that reported in <sup>50</sup>, where the disorder dataset corresponded to sequence segments as predicted by Mobi-DB <sup>51</sup> on the DisProt v7.0 sequences. The *pyHCA* GitHub repository has been updated accordingly (<https://github.com/DarkVador-HCA/pyHCA>).

## Sequence datasets

### (a) Sequence redundancy filters

For each dataset described below (b-e), sequence redundancy was addressed using *mmseqs2*<sup>52</sup> clustering (mode 1, sensitivity 8) with a sequence identity threshold of 30% and a coverage threshold of 90%. The non-redundant sets of sequences (**Supplementary Dataset S2**) comprised the representative sequences of each cluster.

### **(b) Disordered segments**

Disordered sequences, as assessed from experiments and manually curated, were extracted from the reference database DisProt v8.0.2 (8 254 sequences,<sup>20</sup>; <https://disprot.org/>). The corresponding non-redundant set DisProt comprises 3 166 sequences.

### **(d) Soluble domains**

27 543 sequences of soluble domains with known 3D structures were collected from the Structural Classification of Proteins — extended (SCOPe) v2.0.7 database (<sup>53</sup>; <https://scop.berkeley.edu/>) as provided by Astral repository with 95% identity filter. The SCOPe classification of these entries according to their content in secondary structures was as follows: 4 974 all-alpha domains (class a), 7 622 all-beta domains (class b), 8 250 alpha/beta domains (class c), 6 697 alpha+beta domains (class d). Our non-redundant dataset SCOPe comprises 10 885 domain sequences with the SCOPe a-d classes represented by 2 507, 2 511, 2 841 and 3 026 entries, respectively.

### **(e) Transmembrane domains**

The Orientations of Proteins in Membranes (OPM) database (<sup>54</sup>; <https://opm.phar.umich.edu/>) was evidenced to include the largest number of membrane proteins with known 3D structures<sup>55</sup>. The OPM entries annotated as transmembrane domains were downloaded on August 30, 2021 and include 3 classes: alpha-helical polytopic or multi-pass (140 superfamilies, 5 381 entries), bitopic or single-pass (69 superfamilies, 1 151 entries) and beta-barrels transmembrane (35 superfamilies, 601 entries) domains. 35 sequences only formed by unknown residues (replaced in OPM by alanines) were suppressed. We performed a first redundancy treatment to suppress repeated sequences (using *mmseqs2* clustering with a 0.95 identity threshold and keeping the longest sequence of the cluster), leaving only 3 051 unique TM domains in total. For each domain, OPM provides the information of the calculated transmembrane (TM) segment boundaries. In order to delineate the whole membrane-spanning domains, we included loops with length smaller than or equal to 30 residues in our sequence dataset. According to the distribution of loop lengths in TM domains and to the minimum size of known globular domains, TM segments separated by more than 30 residues are probable to encompass nested soluble domains<sup>56</sup>. In order not to

include these cases of large loops in our sequence dataset, we only kept the first 15 residues after the first TM segment and the last 15 before the second TM segment. If an extended segment boundary falls inside a hydrophobic cluster, we moved it to include the whole hydrophobic cluster and the 4 following residues (*i.e.* the minimal distance considered to separate contiguous hydrophobic clusters, see **Supplementary Dataset S1**). This trimming needed to be applied to 1 882 sequences (see **Supplementary Dataset S2** for details). The non-redundant dataset OPM comprises 1 698 sequences: 1330, 165, and 203 annotated as alpha-helical polytopic, bitopic and beta-barrels transmembrane domains, respectively.

#### **(f) Proteomes from AlphaFold Protein Structure database v1**

The amino acid sequences, the 3D structure predictions and the corresponding per-residue model confidence values (pLDDT) were downloaded from the AlphaFold Protein Structure database (AFDB) v1 (<sup>47</sup>, <https://alphafold.ebi.ac.uk>, downloaded on July 21, 2021) for the reference proteomes of 21 model organisms. pLDDT values estimate on a per residue basis how well the predicted structure would agree with the experimental 3D structure and is scaled between 0 and 100 as follows: very low (pLDDT  $\leq$  50), low ( $50 < \text{pLDDT} \leq 70$ ), confident ( $70 < \text{pLDDT} \leq 90$ ), very high (pLDDT  $> 90$ ).

#### **Figure creation and statistical analyses**

3D structures were visualized with the UCSF Chimera package <sup>57</sup>. Statistical analyses were performed using the R software, version 4.1.2 (<sup>58</sup>) and the Python Language (<http://www.python.org>), versions 3.7.6 (for the HCA score implementation and AFDB v1 analyses) and 3.6.3 (for the classification of order and disorder from DisProt, SCOPe and OPM databases). Standardized principal component analysis (PCA) and hierarchical clustering on principal components were performed using the R package Factoshiny, version 2.4 (<https://cran.r-project.org/web/packages/Factoshiny>). Mean comparisons by non-parametric Mann-Whitney U test were performed using the Python scipy library, version 1.4.1 (<https://scipy.org>). Graphics were generated using the Python libraries matplotlib, version 3.0.3 (<https://matplotlib.org>) and seaborn, version 0.11.2 (<https://seaborn.pydata.org>).

## RESULTS

### Exploring the degree of order and disorder in reference databases using HCA score

HCA score was introduced in order to provide a global estimation of the density of a given amino acid sequence in hydrophobic clusters and hydrophobic amino acids within hydrophobic clusters, as a proxy for order/disorder ratio in protein segments. Hence, we focused on foldable segments (as defined by the *segment* function of the *pyHCA* package) of the reference datasets and optimized the weights of this metrics to best separate the different categories according to structural features.

We optimized the weights of this metrics for separating at best order and disorder relative to two datasets: (i) a set of 16489 non-redundant foldable segment sequences from globular proteins from the reference databases SCOPe for soluble domains (14624 segments) and OPM for transmembrane domains (1865 segments) and (ii) a set of 3276 non-redundant foldable segment sequences from disordered proteins from the reference DisProt database (v8.0.2) (**Supplementary Dataset S3**). The optimal HCA score was reached for a minimum overlap of 49% between the score distributions of the globular and disordered sequences, respectively, and with the weights of 9 for strong hydrophobic residues in hydrophobic clusters, 7 for non-strong hydrophobic residues in hydrophobic clusters and -10 for residues outside of hydrophobic clusters (see equation (1) in the Material and Methods section).

A vast majority of amino acids of the SCOPe and OPM non-redundant datasets are included in foldable segments (90 % and 96 %, respectively), while non foldable segments in these datasets mostly correspond to large, hydrophilic loops (see for instance the example shown in **Supplementary Figure S1a**). In contrast, only 50% of amino acids in the DisProt v8.0.2 dataset are included in foldable segments, the remaining ones (*i.e.* in non-foldable segments) can be thus considered as fully disordered (**Supplementary Figure S1b**). DisProt segments can thus be analyzed after being separated into two distinct categories (*i.e.* foldable and non-foldable segments). A few instances of poor coverage of folded domains of the SCOPe database by foldable segments are observed. These correspond to sequences rich in alanine (class a) and threonine/serine (class b), which are not included in the hydrophobic alphabet used for hydrophobic cluster definition, but participate in the hydrophobic core (**Supplementary Figure S1c and S1d**). Some other cases of low coverage actually correspond to domains stabilized by a ligand or for which the folding is cooperative or dependent of an oligomeric organization (obligate oligomers) (**Supplementary Figure S1e**).

The HCA score values range between -7 and 9 (**Figure 1, Supplementary Dataset S3**). Soluble domains have HCA scores included mostly in the [-1,3.5] interval (labeled b). As illustrated for the SCOPe b

class, the lowest and highest values are mainly associated with the presence and absence of large loops (thus large inter-cluster linkers), respectively (**Figure 2a**). HCA scores for the beta-barrel transmembrane class are close to those of this SCOPe distribution, while those for the alpha-helical polytopic class are higher ([3.5,6.7] interval, (labeled c)) (**Figure 1**). This is consistent with the general properties of these two classes of proteins; the beta-barrel strands, albeit longer, are indeed characterized by a periodicity of 2 in hydrophobic amino acids, as for beta-strands from globular domains, while transmembrane alpha-helices from alpha-helical polytopic segments are associated with larger numbers of strong hydrophobic amino acids dotted with charged/polar amino acids, as well as tiny ones (Gly, Ala), ensuring tight packing (**Figure 2b**). Beta-barrel transmembrane sequences can be distinguished from soluble globular domains by their amino acid composition and hydrophobic cluster characteristics (**Figure 2b**, also see Discussion). Finally, a separate category exists for all reference databases, with HCA scores above a value of 7.6 (labeled d), including mostly short sequences covered by a single hydrophobic cluster. These correspond entirely to the foldable segments of OPM bitopic sequences (made of only one transmembrane segment, rich in strong hydrophobic amino acids, **Figure 2b** light blue) or are part of OPM polytopic or SCOPe soluble sequences, which have been fragmented into several parts due to the presence of large loops (see the example in **Supplementary Figure S1a**). Overall, the HCA score represents a metric that allows to globally assess the order/disorder ratio in a single protein sequence and to unravel its main structural features.

The HCA score distribution of foldable segments from the DisProt database is wide, ranging from -7 to 9, due to the structural heterogeneity of the corresponding sequences, reflecting diverse order/disorder ratio. Some foldable segments have low HCA score values, having hydrophobic cluster densities below those observed for folded domains, which reflect their propensity to fluctuate between disorder and order and/or to interact with partners (labeled a and a' in **Figure 1**). This is for instance the case of a foldable segment from the yeast nuclear pore complex NUP2 protein, which contains FxFG repeats (shown with a red circle on **Figure 3a**, HCA score of -5.36 (*a* region in **Figure 1**)). FxFG repeats are known to bind transport factors of the importin-beta/karyopherin-beta family, which function as carriers for many nuclear trafficking processes<sup>59</sup>. They bind in hydrophobic pockets displayed at the surface of HEAT repeats, in which the phenylalanine side chains are buried (modeled on **Figure 3a**, on the basis of a complex of FxFG peptides with importin-beta). A second example is related to foldable segments of the mouse nucleoporin ELYS, interacting with chromatin<sup>60</sup>(**Figure 3b**, no 3D structure available, HCA score of -2.84 (*a'* region in **Figure 1**)). A third example is that of a cadherin 1 segment, whose 3D

structure has been partially solved in complex with catenin beta 1<sup>61</sup> (**Figure 3c**, HCA score of -1.62 (*a'* region in **Figure 1**)). Some other DisProt foldable segments are falling into the “folded, soluble domain” category (*b* region in **Figure 1**), being characterized by hydrophobic clusters ratio typical of this kind of stable 3D structures. This is for instance the case of a foldable segment of the 7SK Sn RNA methylphosphate capping enzyme, which is partially disordered and undergoes a disorder-to-order conformational change upon RNA binding<sup>62</sup> (**Figure 3d**, HCA score of 0.40). Another example of a DisProt sequence, totally covered by a foldable segment, is shown in **Supplementary Figure S2a**, with a cluster composition and density similar to that observed in globular domains, albeit with a slightly low content (30%) in hydrophobic amino acids. This example corresponds to the yeast proteasome maturation factor Ump1, which is disordered when free in solution, as observed using various experimental techniques<sup>63,64</sup>. It however forms well-ordered secondary structures in complex with the 20S core particle, playing a key role in the dynamical assembly of proteasome<sup>65</sup>. An additional example of stabilization of 3D structures in complexes is provided in **Supplementary Figure S2b**. It is also in this region *b*, typical of well folded 3D structures, that are found molten globules (IDP0:00077), which are compact, with native-like secondary structures but disordered tertiary structures<sup>10</sup>. We illustrate this case with the well-known nuclear coactivator binding domain (NCBD) of the mouse CREB-binding protein (DisProt DP00348r018 aa 2059-2117; foldable segment aa 2068-2102, HCA score = -0.09)<sup>66</sup>, which moreover folds into two different conformations depending on the binding partner<sup>67</sup> (**Supplementary Figure S2c**).

Next, higher HCA scores values, typical of polytopic membrane domains (*c* region in **Figure 1**), are also encountered in DisProt sequences. In a general way, these segments have long hydrophobic clusters (which are common in membrane domains), but form long helical structures involved in oligomeric coiled-coils. This is the case for instance of the HR2 domain of SARS-CoV-2 spike glycoprotein, which form an elongated six-helix bundle together with the HR1 domain<sup>68</sup> (**Figure 3e**, HCA score of 5.66). Another example is the flexible C-terminal part of the *E. coli* antitoxin ParD, which is involved in the binding and neutralization of the ParE toxin<sup>69</sup> and forms upon binding long helices docking into the groove of the partner<sup>70</sup> (DP0033r012, HCA score:5.15, 27 amino acids).

Finally, HCA score values above 7.6 (*d* region in **Figure 1**) include foldable segments with a single hydrophobic cluster, as for similar segments of the SCOPe and OPM databases. Such segments frequently correspond to preformed secondary structures (MORFs) or short linear motifs (SLIMs), which fold upon binding, as illustrated here with the short linear motif of the *C. elegans* EGL-1, whose binding to CED-9 initiate cell programmed death<sup>71</sup> (**Figure 3F**, HCA score of 8.23).



Overall, applied to disordered sequences extracted from the DisProt database, the HCA score also allows to globally assess the order/disorder ratio and unravel the wide diversity (or different flavors) of disorder.

### Leveraging AlphaFold2 predictions with *pyHCA* package

The recent development of the AlphaFold2 (AF2) deep-learning approach was a huge step forward in the high-throughput prediction of 3D structures for folded regions, and was even announced as a disorder prediction tool, as region with very low confidence largely overlap with IDRs<sup>30-32,72</sup>. We wanted here to compare the AF2 predictions to those of foldability and structural states, which can be made using *pyHCA*. Each proteome processed by AF2, provided by AlphaFold Protein Structure Database (AFDB<sup>47</sup>, <https://alphafold.ebi.ac.uk>), can be described in terms of foldability and order/disorder content by the *pyHCA* package: the *segment* function allows first to quantify the proteome coverage in foldable segments, then to compute the HCA score on foldable segments, leading to their assignment to a structural class related to their position within the order-disorder continuum, as inferred in **Figure 1**. The link between these two metrics (one, binary, and the other, discretized from a continuum) and the per-residue metrics of uncertainty (pLDDT) defined by AF2 was here explored. Four main classes of pLDDT values, reflecting the confidence in the AF2 structural predictions, are generally considered: very high (pLDDT > 90), high ( $90 \geq \text{pLDDT} > 70$ ), low ( $70 \geq \text{pLDDT} > 50$ ) and very low (pLDDT ≤ 50).

We first analyzed the confidence in the AF2 structural predictions for residues in *versus* outside of the foldable segments, for the 362 094 sequences of the 21 reference proteomes from AFDB v1. Among this whole dataset, 81.9% of the residues are part of foldable segments (ranging from 73.1% to 96.3% for the proteomes of the parasitic protozoa *Leishmania infantum* and the autotrophic hyperthermophilic archaeon *Methanocaldococcus jannaschii*, respectively), corresponding mostly to confident predictions: 45.7% and 27.1% residues with very high and high pLDDT, respectively (**Figure 4a**). Instead, the residues located outside of foldable segments (hereafter described as non-foldable segments, representing 18.1% residues of the whole dataset) correspond mostly to low confident predictions (61.3% and 17.1% residues with very low and low pLDDT, respectively, **Figure 4a**). These trends are also observed for each organism separately, at the exception of *Plasmodium falciparum* whose proteome is dominated by low confident predictions, even in the foldable segments (39.7% and 13.1% residues with very low and low pLDDT, respectively) (**Supplementary Figure S3, Supplementary Table S1**). The four prokaryotic proteomes are instead largely dominated by high confident predictions, in foldable (>72%

residues with very high pLDDT) but also in non-foldable segments (<40% residues with very low or low pLDDT). Overall, low and high pLDDT scores are thus mostly observed for amino acids in non-foldable (full disorder) and foldable segments, respectively (**Figure 4b, Supplementary Table S1**), thereby independently supporting the observation that AF2 low confidence predictions are significantly enriched in intrinsically disordered regions<sup>30,31</sup>. This is consistent with the recently reported distributions of pLDDT scores for the per-residue predictions of order and disorder<sup>72,73</sup>.

In order to further explore the confidence of AF2 predictions for foldable and non-foldable segments separately, we considered regions of contiguous residues (2 aa as minimum length) within these segments where the residues are all affiliated to the same class of pLDDT values (4 classes, see above). In the non-foldable segments, 96.1% residues with a very high pLDDT are located in such regions of homogeneous prediction confidence (3.9% residues with a very high pLDDT in non-foldable segments are thus isolated within regions with a lower prediction confidence).

The sequence length distribution of these regions of very high prediction confidence in non-foldable segments is illustrated in **Supplementary Figure S4** (only 13 sequences with length greater than 100 amino acids are observed, corresponding to 6 different proteomes). These regions fall into two distinct categories, corresponding to: (i) large loops within folded domains (**Supplementary Figure S5a**) or linkers between folded domains (**Supplementary Figure S5b**), which are generally disordered or flexible in isolated proteins but conditionally folded in presence of specific interactions, (ii) regions which are well folded but independent of the presence of strong hydrophobic amino acids (**Supplementary Figure S5c-f**). These sequences correspond to either (a) long coiled-coils, which form extended rod-like structures and are rich in alanine (**Supplementary Figure S5d**) or made of acidic and basic-rich heptad repeats (**Supplementary Figure S5c**), and to various structural repeats (left-handed parallel beta helix repeats (**Supplementary Figure S5e**), tetratricopeptide repeats, armadillo repeats, ...) or (b) domains with ion-dependent folding (calcium (**Supplementary Figure S5f**), zinc, ...). Non-foldable sequences which form well-stable structures thus correspond to particular cases in which the fold is not conditioned by the presence of a core of strong hydrophobic amino acids, but which possess clear, distinctive amino acid composition. Conversely, as discussed in<sup>73</sup>, foldable segments do include sequences with disorder potential, for which AF2 prediction however capture some structural features formed upon interaction (already depicted in experimental structures).

The methodology developed here allows us to put the AF2 prediction made at the residue level into the context of the foldable segment to which it belongs and of its position in the order/disorder continuum.

We thus calculated the distribution of residues in each pLDDT category as a function of the HCA scores of the foldable segments in which they are included (**Supplementary Dataset S4**). Overall, we observed differences in the relative proportions of each pLDDT category as regards to the structural states (labeled a to d), previously defined based on the HCA score (**Figure 5a**). The highest proportion of residues with very high pLDDT (51.8%) is observed for the foldable segments with an HCA score within ]1;2] interval, corresponding to the typical score of soluble domains (*b* region, as defined in **Figure 1**). Moreover, the residues with very high pLDDT represent 48.3% residues included in foldable segments with a HCA score in ]-1;6], corresponding to the soluble/beta-barrel and alpha-transmembrane domains (*b* and *c* regions). Instead, residues with very low pLDDT dominate the likely disordered foldable segments with a HCA score in ]-7;-4] (74.8%), but also those with a HCA score in ]+7;+9] (63.7%) corresponding to local maxima in the HCA score distribution for DisProt disordered sequences (see **Figure 1**). We should notice the uniformity of the length of the foldable segments with HCA score lower than +7, with a median length of 84 aa, as evidenced by the tight correlation between the number of residues and the number of foldable segments within the corresponding intervals of HCA score (see top graph in **Figure 5a**, with black bars and grey line, respectively). Instead, the foldable segments with HCA score higher than +7 are much shorter, with a median length of 6 aa.

Although these global trends of relationship between pLDDT and HCA score of foldable segments are conserved across the 21 individual proteomes, several differences should be noticed, as illustrated by five representative patterns in **Figure 5b-f**. The 21 proteomes have been clustered according to the principal component analysis (PCA) built of their relative proportion of residues of each pLDDT category for foldable segments of each interval of HCA score values (**Figure 6**). The first principal component explained 56.4% of the total variance and showed that the foldable segments of prokaryotic and eukaryotic proteomes differ for almost all intervals of HCA scores, with higher proportions of very high and very low pLDDT residues, respectively (**Supplementary Figure S6a-d**). At second order and according to the second principal component (explaining only 11.7% of the total variance), the proteome of the archaeon *Methanocaldococcus jannaschii* slightly differs from the 3 bacterial proteomes, in particular due to the absence of sequences with HCA scores lower than -4 (corresponding to full disorder) in *M. jannaschii* (**Figure 5b-c**).

The proteomes of eukaryotes are split into 3 groups, cluster 4 includes the four plant proteomes (*Arabidopsis thaliana*, *Oryza sativa*, *Glycine max*, *Zea mays*), the three *Ascomycota* are separated according to their taxonomy: the two *Saccharomycetaceae* (*Candida albicans* and *Saccharomyces*

*cerevisiae*) belongs to cluster 3 whereas *Schizosaccharomyces pombe* is in cluster 4. Cluster 4 is characterized by a higher proportion of low and very low confidence regions in segments with HCA score in ]-1;6] (corresponding to globular sequences), and by a lower proportion of very low confidence predictions in segments corresponding to disorder (HCA score lower than -5) (**Figure 5d-e**). Finally, the proteome of *Plasmodium falciparum* forms another last group, being characterized by predictions of globally lower confidence (**Figure 5f**). This is consistent with the high level of dark sequences in the Apicomplexa proteomes, for which remote homologs cannot easily be detected for efficient covariation analyses.

In the foldable sequences, we observed two major types of apparent discrepancies between the confidence score of AF2 structural prediction and our classification of structural states based on HCA score. These types indeed deviate from the current assumption that the structure of folded regions would be predicted with high confidence, while disordered regions would not. First, we reported cases of regions of contiguous amino acids with high pLDDT values with low HCA scores, that we suggested to correspond to disorder (785 sequences with length greater than 100 amino acids, corresponding to 20 different proteomes, *i.e.* at least one sequence in all AFDB v1 proteomes except *Methanocaldococcus jannaschii*; **Supplementary Figure S7a**). These regions include the two categories depicted before for non-foldable segments with high pLDDT values, *i.e.* (a) disordered sequences included in large loops and/or undergoing conditional folding (AF2 then capturing the structure of the complexed state), (b) long, repetitive structure with a particular abundance of (alpha-forming) alanine or (beta-forming) serine/threonine and sequences with an ion-dependent folding (**Supplementary Figure S8**). Low HCA scores are also observed in some enzymes, which depend on ions for their catalytic activity (*e.g.* one of the longest regions (224 amino acids) in the human carbonic anhydrase (P00915, HCA score of -0.62)). Apolar but non-strong hydrophobic amino acids are also particularly abundant in these low HCA score regions within standard globular domains, completing the hydrophobic clusters participating in the hydrophobic core.

Second, we reported cases of regions of contiguous amino acids with very low pLDDT values but with HCA scores typical of folded domains (9722 sequences with length greater than 100 amino acids, corresponding to 21 different proteomes; **Supplementary Figure S7b-c**). Only 54 regions of more than 30 amino acids are found in the *E. coli* proteome (**Figure 7**), among which a whole protein reported as a recent and rare case of emerging gene by overprinting in this bacterial species (MbiA<sup>74</sup>; **Figure 7a**), small domains in multidomain proteins (RhsC - **Figure 7b**, YdfE - **Figure 7c**) or a protein segment

(YjfZ - **Figure 7d**). RSSs are more or less well predicted by AF2 although their assembly cannot be validated. Of note however is the case of MbiA for which only a part of the RSS is predicted, while the HCA plot clearly indicates several others, associated with hydrophobic clusters typical of beta-strands. This failure of RSS prediction appears recurring in several AF2 predictions of such regions in the *Plasmodium falciparum* proteome, as illustrated in **Figure 7e**. These results indicate that regions with AF2 very low confidence do not always correspond to disordered regions, in line with studies that have compared AF2 to order/disorder predictors <sup>72,73</sup>.

## DISCUSSION

The remarkable progress that has recently been made in the field of structure prediction <sup>75</sup> owes its success to the use of machine learning approaches and the consideration of pre-existing knowledge at the protein sequence and structure levels. In particular, the wealth of evolutionary information was leveraged to an optimal level to extract key features of inter-residue contacts and distances, which have paved the way to unprecedented levels of accuracy in the prediction <sup>28,76</sup>. Considering evolutionary information was also instrumental in the recent advances made in the field of disorder prediction <sup>27</sup>.

Among the key questions that evolutionary-based methods applied to structure prediction cannot easily address, at least with the expected accuracy, is that of regions lacking known homologous sequences and falling outside family annotations, which constitute the dark proteome <sup>77-79</sup>. A common idea was that the dark proteome is largely made up of IDPs/IDRs, as these are difficult to be characterized by traditional methods of structural biology <sup>18</sup> and moreover harbor less amino acid conservation of their sequences <sup>80</sup>. Contrary to this idea, it was shown that the dark proteome contains an important part of non-disordered sequences, constituting the “unknown unknown” <sup>77,79</sup>. A recent survey of the human dark proteome before and after AF2 development has indicated that only a part of these non-disordered sequences was predicted with good accuracy <sup>81</sup>, thereby supporting the ever-present need to develop tools for better characterizing the foldability potential and structural features from the only information of single amino acid sequences and applicable to any proteome, even the darkest ones. It should be noted here that the foldability is linked to the ability of building blocks (*i.e.* the regular secondary structures) to interact. However, the foldability score does not provide information of how these interact, which is typically addressed by analysis of residue co-variation, when homologs are available.

The (non)foldability of proteins is encoded in their amino acid sequences, with two global physicochemical patterns, the absolute mean charge and the mean hydrophathy, primarily accounting for the differences between two classes<sup>82</sup>. This issue of the foldability potential was addressed here in two steps, based on the consideration of this basic hydrophobic/non-hydrophobic dichotomy, enriched by the information on local structure through the use of a two-dimensional representation of amino acid sequences. Capitalizing on the proven ability of HCA to highlight structural invariants in a context of high evolutionary divergence<sup>82</sup>, we do not explicitly use an hydrophobicity scale especially in order to take into account at the best this dichotomy and the driving role of strong hydrophobic amino acids in the formation of regular secondary structures, regardless of their specific physico-chemical features. The first step of our procedure relies on a binary definition of foldability, with the delineation of homogeneous regions in terms of general properties related to order (foldable regions) or disorder (non-foldable regions). The second step, focusing on the foldable regions, estimates their degree of foldability in a continuum. This combined approach goes thus beyond a binary and per-residue order/disorder dichotomy and is independent of the consideration of a set of homologous sequences. *pyHCA* thus provides a useful information about the position and global characteristics of structurally homogenous segments, *i.e.* the foldable segments, which contain building blocks that are likely to interact and thus goes beyond the local interactions which can be predicted by considering only isolated hydrophobic clusters. *pyHCA* is in line with the spirit of polymer scaling behavior, which combines hydrophathy and charge patterning and is used for characterizing the structural properties of IDPs/IDRs<sup>83</sup>. It is also to be compared to ODiNPred, a sequence order/disorder predictor which uses a deep neural network trained on a database of an experimental, continuous-valued quantification of local disorder based on NMR chemical shifts and considering a large number of sequence features<sup>33</sup>, among which foldable domains as predicted by *SEG-HCA*. The foldable segments defined by *SEG-HCA* are expected to fold spontaneously or conditionally into stable 3D structures through the participation in an hydrophobic core, while non-foldable segments correspond to full disorder, with the exception of regions whose stably fold without the need of a consistent hydrophobic core, but *e.g.* depending on ion binding<sup>49</sup>. Conditional foldable regions, having transient residual structures or fold dependent on interactions or environment<sup>8,11</sup>, can generally be distinguished from the autonomous folding units as these segments are often predicted as disordered by current disorder predictors<sup>42</sup>. This category of transient disorder includes short linear motifs (SLIMs)<sup>84</sup> and Molecular Recognition Features (MoRFs)<sup>85</sup>, which are generally embedded in large disordered regions. Their intermediate behavior can be highlighted using tools such as *ANCHOR*<sup>86</sup> and visualized with *FELLS*, an estimator of latent structures integrating *SEG-HCA* and *IUPred2*

predictions<sup>87</sup>. It is interesting to note that these sequences, which are predicted as disordered but foldable, are globally well predicted by AF2, capturing the folded state, however without detecting their structural plasticity<sup>73</sup>.

In a global way, the HCA scoring scheme introduced here allows to appreciate the degree of foldability of protein segments, reflecting the relative abundance of loop/coil regions (*disorder*) and regular secondary structures (*order*). Thereby, we are able to disentangle the great diversity present within the IDPs/IDRs, which is reflected by a wide range of HCA score values, whereas folded domains are characterized by narrower ranges of values (Figure 1). The distinct behaviors between these two groups have also been evidenced in a recent study using a Gini index, which allows to estimate distribution uniformity<sup>88</sup>.

Based on the HCA score, some foldable segments from DisProt clearly deviate from folded-like regions (regions *a - a'* in **Figure 1**) while having the capacity to conditionally fold. This is the case, for instance, of the segments shown in **Figure 3a to c**, with low HCA scores. These segments can thus be easily distinguished from those that are closer to a soluble, globular domain behavior (region *b* in **Figure 1**). IDPs/IDRs found in this last region are difficult to distinguish from well-folded globular domains based on the only consideration of the HCA score and other features must be considered to refine the analysis. However, one can note that such IDPs/IDRs are generally shorter than typical well-folded globular domains extracted from the SCOPe database, which may explain that they are unable to fold stably in absence of partners. Indeed, only 48.4% of sequences from this category in DisProt have length greater than 30 amino acids (mean length 81.6 aa), against 85.1 % in SCOPe (mean length 133.3 aa). In cases of longer IDPs/IDRs from this category, amino acid composition may help to distinguish them from well-folded globular domains. Indeed, we observed that segments from DisProt of this category (region *b* in **Figure 1**) are enriched in polar amino acids that have been previously described as disorder-promoting (Gln, Lys, Ser, Glu)<sup>89</sup> (**Supplementary Figure S9a**). Remarkably, they have composition in strong hydrophobic amino acids comparable to that found in soluble domains. This composition, consistent with high HCA scores, thus defines a specific class of long disordered sequences, reflecting their propensity to fold upon constraint (**Supplementary Figure S10**). The example of AF4-AF9 complex, discussed in<sup>49</sup> also suggested that some sequences of IDPs/IDRs might be stabilized in absence of interacting partners by intra-molecular interactions mediated by sequences located at long-range distance in the protein. In this region *b*, one can also observe that the composition of well-folded soluble domains is different from that of also well-folded transmembrane beta-barrels, characterized by similar HCA scores values (**Figure**

1). These have indeed distinctive features such as dyad-repeat patterns and a high abundance in aromatic amino acids at the bilayer interface <sup>90</sup>, allowing their accurate predictions by dedicated tools (*e.g.* <sup>91,92</sup>). Here, we also evidenced an enrichment of beta-barrel foldable segments relative to soluble domain ones in Tyr and Trp, as well as in small and polar amino acids (Gly, Asn, Ser, Thr;) (**Supplementary Figure S9a**), consistent with previous observations <sup>90</sup>. In the *c* region, the DisProt foldable domains can also be distinguished of well-folded alpha-helical membrane domains by their amino acid composition, as the former ones are also enriched in polar/charged amino acids (Asp, Glu, Asn, Gln, Arg, Lys, His, Ser, Thr) (**Supplementary Figure S9b**). Examples extracted from this category of DisProt segments highlighted sequences with long hydrophobic clusters (length similar to transmembrane helices), but forming elongated, soluble coiled-coils. First attempts to develop tools for sorting sequences in the *b-c* regions and distinguishing between stable structures and conditional order or molten globules, based on these amino acid composition differences, are encouraging. However, further investigations are needed to characterize the building blocks (hydrophobic clusters) of IDPs and sequences linking them, in order to understand the molecular basis of their particular structural behavior, especially in terms of both fuzziness (typified by the co-existence of several minima of free-energy content) <sup>93,94</sup> and frustration <sup>95,96</sup>.

The combination of HCA score and AF2 pLDDT applied to proteome-wide analysis shed light on the relative part of each given proteome in which order is still hidden, corresponding to very low AF2 pLDDT values but with HCA scores typical of globular-like regions (regions *b - c* in **Figure 1**). Examination of particular cases of hidden order indicates that AF2 is able in some situations to predict RSSs, which correlate with hydrophobic clusters (as observed in the *E. coli* sequences presented in **Figure 7**), nonetheless without confidence in the way they are associated. The overestimation of disorder by AF2 is minimum in the case of prokaryotic proteomes (**Figure 8a**), where large amount of known 3D structures and sequences are available. In contrast, the accuracy of disorder prediction by AF2 is much lower in case of *Dictyostelium discoideum* and parasitic organisms such as *Trypanosoma cruzi*, *Leishmania infantum* and *Plasmodium falciparum*. For this latter proteome, a percentage of amino acids as high as 40.8% identified in the globular-like category by *pyHCA* correspond to AF2 very low pLDDT scores. Some hypothetical proteins from *Plasmodium falciparum* escape RSS prediction by AF2 and are represented as fully disordered, although they possess hydrophobic clusters with HCA scores typical of well-folded domains. In fact, the higher proportion of amino acids with very low pLDDT values in the *Plasmodium falciparum* proteome (46.0%) cannot be explained in a straightforward way by a higher proportion of disorder, but instead by the compositional bias and low complexity regions leading to mask the order characteristics and to leave a large number of sequences in the dark <sup>97,98</sup>. *pyHCA* is not affected



by these biases and estimates a similar foldability trend for *Plasmodium falciparum* as for other eukaryotes such as human (80-85% aa within foldable segments, **Table S1**). It allows therefore to unravel the characteristics of the hidden order, as already applied to the identification of hidden actors of the transcription machinery <sup>99</sup>.

Importantly, the reliability of the “order” prediction by pyHCA can unambiguously be demonstrated in cases where the AF2 prediction fails. Indeed, in a joint study <sup>100</sup>, we have focused on long (> 30 amino acids) foldable segments with HCA scores typical of soluble, globular-like domains and that, in the corresponding AF2 models, include only amino acids very low pLDDT values and lack regular secondary structures and folded hydrophobic core (random coils). We showed that some of these sequences are reported in the DisProt database as forming transient regular secondary structures (as determined by various biophysical approaches) <sup>100</sup>. Moreover, a specific search of these segments against the PDB sequences that were published after the deposit of the AF2 models highlighted a case of an AF2 structureless coil prediction, which was further investigated by NMR and ESR, combined with Rosetta computations <sup>101</sup>. In this study, the observed 3D structure of the intracellular domain (ICD) of human alpha7 nicotinic acetylcholine receptor, which shows conformational plasticity, was shown to be well organized in its resting state (**Supplementary Figure S11**). In particular, a loop is anchored onto the intracellular MA helix (merging with the transmembrane helix TM4) via an alpha-helix (h3), which provides significant hydrophobic contributions to the packing of the MA bundle. The foldable segment contains three hydrophobic clusters, one of which corresponding to the observed helix h3. Thus, these examples of conditional order illustrate the relevance of *pyHCA* for identifying foldability potential in the AF2 structureless coil predictions. This opens the way to reveal novel 3D structures, including corresponding to unconditional order, which cannot be predicted by AF2, due to lack of homologs and few local patterns corresponding to existing structures.

*pyHCA* biases mainly rely on regions which stably fold without the need of a consistent hydrophobic core (**Supplementary Figure S5**). The analysis of AF2 pLDDT scores outside of foldable segments provides therefore a useful way to evaluate the overestimation of full disorder by *pyHCA* (**Figure 8b**). At the proteome scale, this bias is minimum for *Dictyostelium discoideum* and the three parasitic organisms (in case of *Plasmodium falciparum*, 3.3% aa outside of foldable segments correspond to a very high pLDDT in AF2 predictions) while ranging to a maximum for the four prokaryotic organisms (51% in *E. coli*). Reminding that these latter proteomes are mostly covered by regions identified as foldable by *pyHCA* (84-96% aa), this overestimation of disorder by *pyHCA* is however quite low when

considering the number of long regions (length higher than 30 aa) capable to fold (*e.g.* corresponding only to 8 different proteins for *Methanocaldococcus jannaschii*).

Overall, combining *pyHCA* with AF2 provides a revised estimation of the full disorder content in proteomes, corresponding to not only non-foldable segments which are not well-predicted by AF2, but also foldable segments with HCA score lower than -4.7. The conditionally folded regions, corresponding to foldable segments with higher HCA score values, are thus discarded from this estimation. Compared to previous studies (*e.g.* <sup>102</sup>), we thus suggest a lower disorder content, as long IDRs (length higher than 30 aa) were found in 12 to 40.4% of long proteins (length higher than 60 aa) in the eukaryotic proteomes in AFDB, and in 0.6 to 4.7% for the prokaryotic ones. Our approach also allows to explore the order/disorder content of proteomes in relation to organism ecological traits, in a complementary way to previous works <sup>24,103</sup>. In addition to a specific behavior of parasitic organisms within eukaryotes (see above and **Figure 6**) that remains to be explained, our study also detected a particularly low disorder content in the proteome of the hyperthermophilic archaeon *Methanocaldococcus jannaschii*, likely related to the high thermal stability constrained by the environmental conditions.

Finally, the scoring scheme offered by *pyHCA* can be used for understanding the protein evolutionary trajectories at the proteome level. It is particularly well-suited to study the structural properties of proteins encoded by *de novo* emerging genes and how such properties have influenced their early emergence and long-term retention. In particular, it can bring new light to the debate of whether *de novo* proteins have much intrinsic disorder <sup>104</sup> or are aggregation-prone <sup>105</sup> and whether retention of *de novo* gene precursor is driven by such properties or remains a stochastic process <sup>106,107</sup>. Several works have already used the concept of foldable segments to investigate such properties <sup>108-110</sup>, and a recent work considering a preliminary version of HCA scoring systems has evidenced that most yeast intergenic ORFs contain the elementary building blocks of protein structures <sup>50</sup>. The example of *E. coli* MbiA, an overlapping (protein-coding) orphan gene which has recently evolved by overprinting and was shown to share the structural properties of globular-like domains although not predicted by AF2 (**Figure 7a**), well illustrates the interest of our approach to decipher structural features in absence of homologs and uncover dark sides of protein evolution.

#### DATA AVAILABILITY

All data and scripts used in the present work are available in a GitHub repository at

<https://github.com/DarkVador-HCA/Order-Disorder-continuum>. The package *pyHCA* is provided at <https://github.com/DarkVador-HCA/pyHCA> under the CeCILL-C license agreement.

## FUNDING

This work has been supported by the Agence Nationale de la Recherche (grant number ANR-14-CE10-0021 and ANR-17-CE12-0016) and the Institut National du Cancer (grant number PLBIO14-299). AB was supported by the PhD program of Doctoral School "Complexité du Vivant" (ED515, Sorbonne Université). Analyses were processed with the support of the computer cluster "Plateforme Calcul Intensif Algorithmique" (UMS2700-PCIA) of the Muséum National d'Histoire Naturelle (MNHN). The authors thank Jean-Paul Mornon for critical reading of the manuscript.

**CONFLICT OF INTEREST STATEMENT:** none declared.

## REFERENCES

1. Kolodny R, Pereyaslavets L, Samson AO, Levitt M. On the universe of protein folds. *Annual Rev Biophys.* 2013; 42:559-582.
2. Nepomnyachiy S, Ben-Tal N, Kolodny R. Global view of the protein universe. *Proc Natl Acad Sci USA.* 2014;111:11691–11696.
3. Han X, Sit A, Christoffer C, Chen S, Kihara D. A global map of the protein shape universe. *PLoS Comput Biol.* 2019;15:e1006969.
4. Schaeffer RD, Kinch LN, Pei J, Medvedev KE, Grishin NV. Completeness and Consistency in Structural Domain Classifications. *ACS Omega.* 2021;6:15698-15707.
5. Oldfield CJ, Dunker AK. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem.* 2014;83:553–584.
6. van der Lee R, Buljan M, Lang B, et al. Classification of intrinsically disordered regions and proteins. *Chem Rev.* 2014;114:6589-6631.
7. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol.* 2015;16:18-22.
8. Uversky VN. Functional roles of transiently and intrinsically disordered regions within proteins. *FEBS J.* 2015;287:1182–1189
9. Uversky VN. Intrinsically Disordered Proteins and Their “Mysterious” (Meta)Physics. *Front Phys.* 2019;7:10.
10. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol.* 2005;6(3):197-208.

## References

11. Jakob U, Kriwacki R, Uversky VN. Conditionally and transiently disordered proteins: Awakening cryptic disorder to regulate protein function. *Chem Rev.* 2014;114:6779–6805.
12. Banani SF, Lee HO, Hyman AA, Rosen MK. Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol.* 2017;18:285-298.
13. Wu H, Fuxreiter M. The structure and dynamics of higher-order assemblies: amyloids, signalosomes and granules. *Cell.* 2016;165:1055-1066.
14. Tompa P, Fuxreiter M. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends Biochem Sci.* 2008;33:2-8.
15. Lyle N, Das RK, Pappu RV. A quantitative measure for protein conformational heterogeneity. *J Chem Phys.* 2013;139:121907.
16. Wright PE, Dyson HJ. Linking folding and binding. *Curr Opin Struct Biol.* 2009 19:31-38.
17. Borgia A, Borgia MB, Bugge K, et al. Extreme disorder in an ultrahigh-affinity protein complex. *Nature.* 2018;555:61-66.
18. Bhowmick A, Brookes DH, Yost SR, et al. Finding Our Way in the Dark Proteome. *J Am Chem Soc.* 2016;138:9730-9742.
19. Lieutaud P, Ferron F, Uversky AV, Kurgan L, Uversky VN, Longhi S. How disordered is my protein and what is its disorder for? A guide through the “dark side” of the protein universe. *Intrinsically Disord Proteins.* 2016;4:e1259708.
20. Hatos A, Hajdu-Soltész B, Monzon AM, et al. DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* 2020;48:D269–D276.
21. Dosztányi Z. Prediction of protein disorder based on IUPred. *Protein Sci.* 2018 27:331-340.
22. Orlando G, Raimondi D, Codicè F, Tabaro F, Vranken W. Prediction of Disordered Regions in Proteins with Recurrent Neural Networks and Protein Dynamics. *J Mol Biol.* 2022;434:167579.
23. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol.* 2004;337:635-644.
24. Xue B, Dunker AK, Uversky VN. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J Biomol Struct Dyn.* 2012;30:137-149.
25. Peng Z, Yan J, Fan X, et al. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell Mol Life Sci.* 2015;72:137-151.
26. Oates ME, Romero P, Ishida T, et al. D2P2: database of disordered protein predictions. *Nucl Acids Res.* 2012;41(D1):D508-D516.
27. Necci M, Piovesan D, Predictors C, Curators D, Tosatto SCE. Critical assessment of protein intrinsic disorder prediction. *Nat Methods.* 2021;18:472-481.
28. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596:583–589.
29. Tunyasuvunakool K, Adler J, Wu Z, et al. Highly accurate protein structure prediction for the human proteome. *Nature.* 2021 596:590-596.
30. Akdel M, Pires DEV, Porta Pardo E, et al. A structural biology community assessment of AlphaFold 2 applications. *bioRxiv.* 2021:2021.2009.2026.461876.
31. Wilson CJ, Choy WY, Karttunen M. AlphaFold2: A Role for Disordered Protein/Region Prediction? *Int J Mol Sci.* 2022;23(9):4591.
32. Ruff KM, Pappu RV. AlphaFold and Implications for Intrinsically Disordered Proteins. *J Mol Biol* 2021;433:167208.
33. Dass R, Mulder FAA, Nielsen JT. ODINPred: comprehensive prediction of protein order and disorder. *Sci Rep.* 2020;10:14780.
34. Zhang T, Faraggi E, Li Z, Zhou Y. Intrinsically semi-disordered state and its role in induced folding and protein aggregation. *Cell Biochem Biophys.* 2013;67(3):1193-1205.

35. Callebaut I, Labesse G, Durand P, et al. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol Life Sci.* 1997;53:621-645.
36. Eudes R, Le Tuan K, Delettré J, Mornon J-P, Callebaut I. A generalized analysis of hydrophobic and loop clusters within globular protein sequences. *BMC Struct Biol.* 2007;7:2.
37. Gaboriaud C, Bissery V, Benchetrit T, Mornon J-P. Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. *FEBS Lett.* 1987;224(1):149–155.
38. Lamiable A, Bitard-Feildel T, Rebehmed J, et al. A topology-based investigation of protein interaction sites using Hydrophobic Cluster Analysis. *Biochimie.* 2019;167:68-80.
39. Woodcock S, Mornon J-P, Henrissat B. Detection of secondary structure elements in proteins by hydrophobic cluster analysis. *Protein Eng.* 1992;5:629–635.
40. Rebehmed J, Quintus F, Mornon JP, Callebaut I. The respective roles of polar/nonpolar binary patterns and amino acid composition in protein regular secondary structures explored exhaustively using hydrophobic cluster analysis. *Proteins.* 2016;84(5):624-638.
41. Bitard-Feildel T, Lamiable A, Mornon J-P, Callebaut I. Order in disorder as observed by the "Hydrophobic Cluster Analysis" of protein sequences. *Proteomics.* 2018;18:e1800054.
42. Faure G, Callebaut I. Comprehensive repertoire of foldable regions within whole genomes. *PLOS Comput Biol.* 2013; 9(10):e1003280.
43. Toth-Petroczy A, Palmedo P, Ingraham J, et al. Structured states of disordered proteins from genomic sequences. *Cell.* 2016;167:158-170.
44. Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* 2003;31:3701-3708.
45. Dosztányi Z, Csizmók V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol.* 2005;347:827-839.
46. Dosztányi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics.* 2005;21:3433–3434.
47. Varadi M, Anyango S, Deshpande M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 2022;50:D439-D444.
48. Faure G, Callebaut I. Identification of hidden relationships from the coupling of hydrophobic cluster analysis and domain architecture information. *Bioinformatics.* 2013;29:1726-1733.
49. Bitard-Feildel T, Callebaut I. HCAtk and pyHCA: A Toolkit and Python API for the Hydrophobic Cluster Analysis of Protein Sequences. *bioRxiv.* 2018:249995.
50. Papadopoulos C, Callebaut I, Gelly JC, et al. Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution. *Genome Res.* 2021;31:2303–2315.
51. Piovesan D, Necci M, Escobedo N, et al. MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.* 2021 49:D361-D367.
52. Steinegger M, Söding J. MMseqs2: sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol.* 2017;35:1026–1028.
53. Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* 2013;42(D1):D304-D309.
54. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.* 2012;40:D370–D376.

## References

55. Shimizu K, Cao W, Saad G, Shoji M, Terada T. Comparative analysis of membrane protein structure databases. *Biochim Biophys Acta Biomembr.* 2018;1860:1077-1091.
56. Ellgaard L, Riek R, Herrmann T, et al. NMR structure of the calreticulin P-domain. *Proc Natl Acad Sci USA.* 2001;98:3133-3138.
57. Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem.* 2004;25:1605-1612.
58. *R: A Language and Environment for Statistical Computing* [computer program]. Vienna: R Core Team; 2021.
59. Denning DP, Patel SS, Uversky V, Fink AL, Rexach M. Disorder in the nuclear pore complex: the FG repeat regions of nucleoporins are natively unfolded. *Proc Natl Acad Sci U S A.* 2003;100(5):2450-2455.
60. Bilokapic S, Schwartz TU. Structural and functional studies of the 252 kDa nucleoporin ELYS reveal distinct roles for its three tethered domains. *Structure (London, England : 1993).* 2013;21(4):572-580.
61. Huber AH, Weis WI. The structure of the beta-catenin/E-cadherin complex and the molecular basis of diverse ligand recognition by beta-catenin. *Cell.* 2001;105(3):391-402.
62. Yang Y, Eichhorn CD, Wang Y, Cascio D, Feigon J. Structural basis of 7SK RNA 5'- $\gamma$ -phosphate methylation and retention by MePCE. *Nat Chem Biol.* 2019;15(2):132-140.
63. Sá-Moura B, Simões AM, Fraga J, et al. Biochemical and biophysical characterization of recombinant yeast proteasome maturation factor ump1. *Comput Struct Biotechnol J.* 2013;7 e201304006.
64. Uekusa Y, Okawa K, Yagi-Utsumi M, et al. Backbone  $^1\text{H}$ ,  $^{13}\text{C}$  and  $^{15}\text{N}$  assignments of yeast Ump1, an intrinsically disordered protein that functions as a proteasome assembly chaperone. *Biomol NMR Assign.* 2014;8:383-386.
65. Schnell HM, Walsh RMJ, Rawson S, et al. Structures of chaperone-associated assembly intermediates reveal coordinated mechanisms of proteasome biogenesis. *Nat Struct Mol Biol.* 2021;28:418-425.
66. Kjaergaard M, Teilum K, Poulsen FM. Conformational selection in the molten globule state of the nuclear coactivator binding domain of CBP. *Proc Natl Acad Sci USA.* 2010;107(28):12535-12540.
67. Kjaergaard M, Andersen L, Nielsen LD, Teilum K. A folded excited state of ligand-free nuclear coactivator binding domain (NCBD) underlies plasticity in ligand recognition. *Biochemistry.* 2013;52(10):1686-1693.
68. Xia S, Liu M, Wang C, et al. Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. *Cell Res.* 2020;30(4):343-355.
69. Oberer M, Zangger K, Gruber K, Keller W. The solution structure of ParD, the antidote of the ParDE toxin antitoxin module, provides the structural basis for DNA and toxin binding. *Protein Sci.* 2007;16(8):1676-1688.
70. Loris R, Garcia-Pino A. Disorder- and Dynamics-Based Regulatory Mechanisms in Toxin-Antitoxin Modules. *Chem Rev.* 2014;114(13):6933-6947.
71. Yan N, Gu L, Kokel D, et al. Structural, biochemical, and functional analyses of CED-9 recognition by the proapoptotic proteins EGL-1 and CED-4. *Mol Cell.* 2004;15(6):999-1006.
72. Aderinwale T, Bharadwaj V, Christoffer C, et al. Real-time structure search and structure classification for AlphaFold protein models. *Commun Biol.* 2022;5(1):316.
73. Alderson TR, Pritišanac I, Moses AM, Forman-Kay JD. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *bioRxiv.* 2022:2022.2002.2018.481080.

74. Fellner L, Bechtel N, Witting MA, et al. Phenotype of htgA (mbiA), a recently evolved orphan gene of Escherichia coli and Shigella, completely overlapping in antisense to yaaW. *FEMS Microbiol Lett.* 2014;350:57-64.
75. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins: Structure, Function, and Bioinformatics.* 2021;89(12):1607-1617.
76. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker DI. Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci USA.* 2020;117:1496-1503.
77. Bitard-Feildel T, Callebaut I. Exploring the dark foldable proteome by considering hydrophobic amino acids topology. *Sci Rep.* 2017;7:41425.
78. Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in 2021. *Nucl Acids Res.* 2021;49:D412–D419.
79. Perdigão N, Heinrich J, Stolte C, et al. Unexpected features of the dark proteome. *Proc Natl Acad Sci USA.* 2015;112(52):15898-15903.
80. Chen JW, Romero P, Uversky VN, Dunker AK. Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. *J Proteome Res.* 2006;5(4):879-887.
81. Porta-Pardo E, Ruiz-Serra V, Valentini S, Valencia A. The structural coverage of the human proteome before and after AlphaFold. *PLoS Comput Biol.* 2022;18:e1009818.
82. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins.* 2000;41(3):415-427.
83. Zheng W, Dignon G, Brown M, Kim YC, Mittal J. Hydropathy Patterning Complements Charge Patterning to Describe Conformational Preferences of Disordered Proteins. *J Phys Chem Lett.* 2020;11(9):3408-3415.
84. Weatheritt RJ, Luck K, Petsalaki E, Davey NE, Gibson TJ. The identification of short linear motif-mediated interfaces within the human interactome. *Bioinformatics.* 2012;28:976–982.
85. Mohan A, Oldfield CJ, Radivojac P, et al. Analysis of molecular recognition features (MoRFs). *J Mol Biol.* 2006;362:1043–1059.
86. Dosztányi Z, Mészáros B, Simon I. ANCHOR: Web server for predicting protein binding regions in disordered proteins. *Bioinformatics.* 2009; 25:2745–2746.
87. Piovesan D, Walsh I, Minervini G, Tosatto SCE. FIELDS: fast estimator of latent local structure. *Bioinformatics.* 2017;33:1889–1891.
88. Carugo O. Hydrophobicity diversity in globular and nonglobular proteins measured with the Gini index. *Protein Eng Des Sel.* 2017;30(12):781-784.
89. Uversky VN, Dunker AK. Understanding protein non-folding. *Biochim Biophys Acta.* 2010;1804(6):1231-1264.
90. Wimley WC. The versatile beta-barrel membrane protein. *Curr Opin Struct Biol.* 2003;13:404-411.
91. Hayat S, Peters C, Shu N, Tsirigos KD, Elofsson A. Inclusion of dyad-repeat pattern improves topology prediction of transmembrane  $\beta$ -barrel proteins. *Bioinformatics.* 2016;32:1571-1573.
92. Tian W, Lin M, Tang K, Liang J, Naveed H. High-resolution structure prediction of  $\beta$ -barrel membrane proteins. *Proc Natl Acad Sci USA.* 2018;115:1511-1516.
93. Miskei M, Horvath A, Vendruscolo M, Fuxreiter M. Sequence-Based Prediction of Fuzzy Protein Interactions. *J Mol Biol.* 2020;432(7):2289-2303.
94. Tompa P, Fuxreiter M. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci.* 2008;33(1):2-8.

## References

95. Freiburger MI, Wolynes PG, Ferreiro DU, Fuxreiter M. Frustration in Fuzzy Protein Complexes Leads to Interaction Versatility. *J Phys Chem B*. 2021;125(10):2513-2520.
96. Malagrino F, Diop A, Pagano L, Nardella C, Toto A, Gianni S. Unveiling induced folding of intrinsically disordered proteins - Protein engineering, frustration and emerging themes. *Curr Opin Struct Biol*. 2022;72:153-160.
97. Pizzi E, Frontalli C. Low-Complexity Regions in Plasmodium falciparum proteins. *Genome Res*. 2001;11:218-229.
98. Hamilton WI, Claessens A, Otto TD, et al. Extreme mutation bias and high AT content in Plasmodium falciparum. *Nucleic Acids Res*. 2017;45:1889–1901.
99. Callebaut I, Prat K, Meurice E, P MJ, Tomavo S. Prediction of the general transcription factors associated with RNA polymerase II in Plasmodium falciparum: conserved features and differences relative to other eukaryotes. *BMC Genomics*. 2005;6:100.
101. Bruley A, Mornon J-P, Duprat E, Callebaut I. Digging into the 3D structure predictions of AlphaFold2 with low confidence: disorder and beyond. *Biomolecules* 2022;12:1467.
100. Bondarenko V, Wells MM, Chen Q, et al. Structures of highly flexible intracellular domain of human  $\alpha 7$  nicotinic acetylcholine receptor. *Nat Comm*. 2022;13(1):793.
101. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform*. 2000;11:161-171.
102. Tang Q-Y, Ren W, Wang J, Kaneko K. The Statistical Trends of Protein Evolution: A Lesson from AlphaFold Database. *Mol Biol Evol*. 2022:msac197.
103. Basile W, Salvatore M, Bassot C, Elofsson A. Why do eukaryotic proteins contain more intrinsically disordered regions? *PLoS Comput Biol*. 2019;15(7):e1007186.
104. Vakirlis N, Acar O, Hsu B, et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat Commun*. 2020;11(1):781.
105. Schmitz JF, Ullrich KK, Bornberg-Bauer E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat Ecol Evol*. 2018;2(10):1626-1632.
106. Dowling D, Schmitz JF, Bornberg-Bauer E. Stochastic Gain and Loss of Novel Transcribed Open Reading Frames in the Human Lineage. *Genome Biol Evol*. 2020;12(11):2183-2195.
107. Grandchamp A, Berk K, Dohmen E, Bornberg-Bauer E. New Genomic Signals Underlying the Emergence of Human Proto-Genes. *Genes*. 2022;13(2):284.
108. Klasberg S, Bitard-Feildel T, Callebaut I, Bornberg-Bauer E. Origins and structural properties of novel and de novo protein domains during insect evolution. *FEBS J*. 2018;285(14):2605-2625.
109. Watson AK, Lopez P, Baptiste E. Hundreds of out-of-frame remodeled gene families in the *Escherichia coli* pangenome. *Mol Biol Evol*. 2021;39(1):msab329.

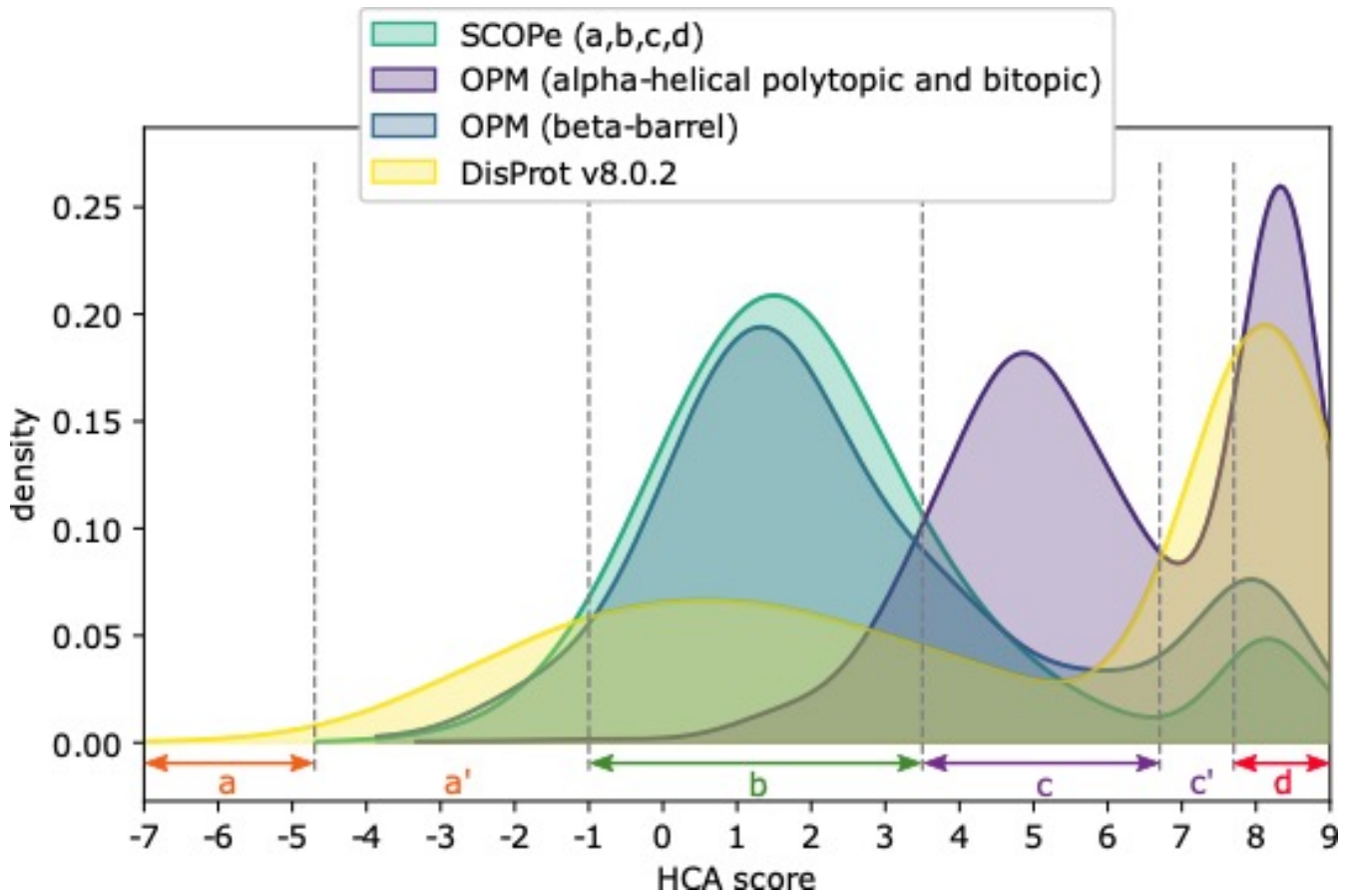


## TABLES AND FIGURES

dataset	class	sequences	aa	sequences with FS	sequences without FS	aa in FS	aa outside of FS
SCOPe	a	2507	347268	2507 (100.0)	0 (0.0)	309100 (89.0)	38168 (11.0)
SCOPe	b	2511	366584	2511 (100.0)	0 (0.0)	322970 (88.1)	43614 (11.9)
SCOPe	c	2841	688643	2841 (100.0)	0 (0.0)	633955 (92.1)	54688 (7.9)
SCOPe	d	3026	455400	3026 (100.0)	0 (0.0)	410347 (90.1)	45053 (9.9)
DisProt v8.0.2		3166	209812	2418 (76.4)	748 (23.6)	105390 (50.2)	104422 (49.8)
OPM	polytopic	1330	186220	1328 (99.8)	2 (0.2)	181396 (97.4)	4824 (2.6)
OPM	bitopic	203	4390	202 (99.5)	1 (0.5)	4025 (91.7)	365 (8.3)
OPM	beta	165	37320	164 (99.4)	1 (0.6)	33986 (91.1)	3334 (8.9)

**Table 1. Foldable segments (FS) in the non-redundant sequence datasets.**

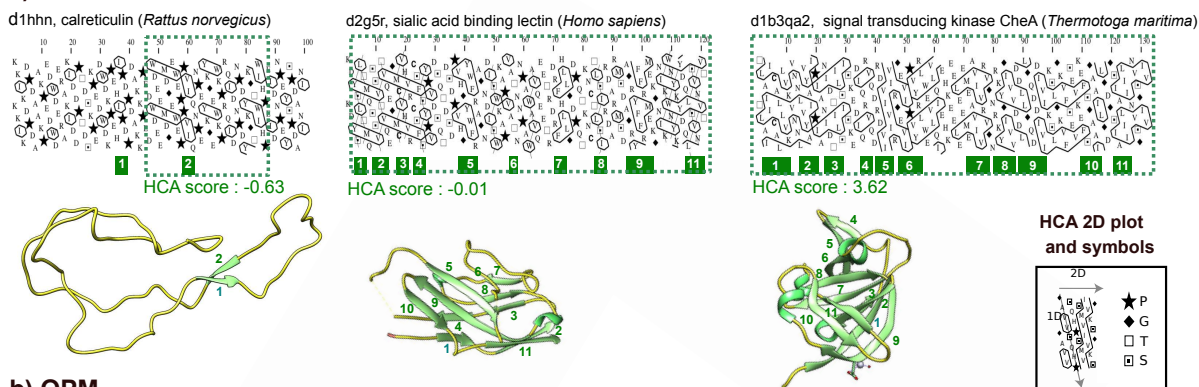
Total number of sequences and amino acids (aa) of each dataset, as well as their number (and percentage) of sequences with and without foldable segment(s), and the number (and percentage) of residues found in or outside of foldable segments. The datasets are those of soluble domains with known 3D structures (SCOPe), transmembrane domains with known 3D structure (OPM) and disordered segments (DisProt v8.0.) (see *Materials and Methods* for details). OPM classes have been shortened to polytopic for alpha-helical polytopic domains, bitopic for alpha-helical bitopic domains and beta for beta-barrels.



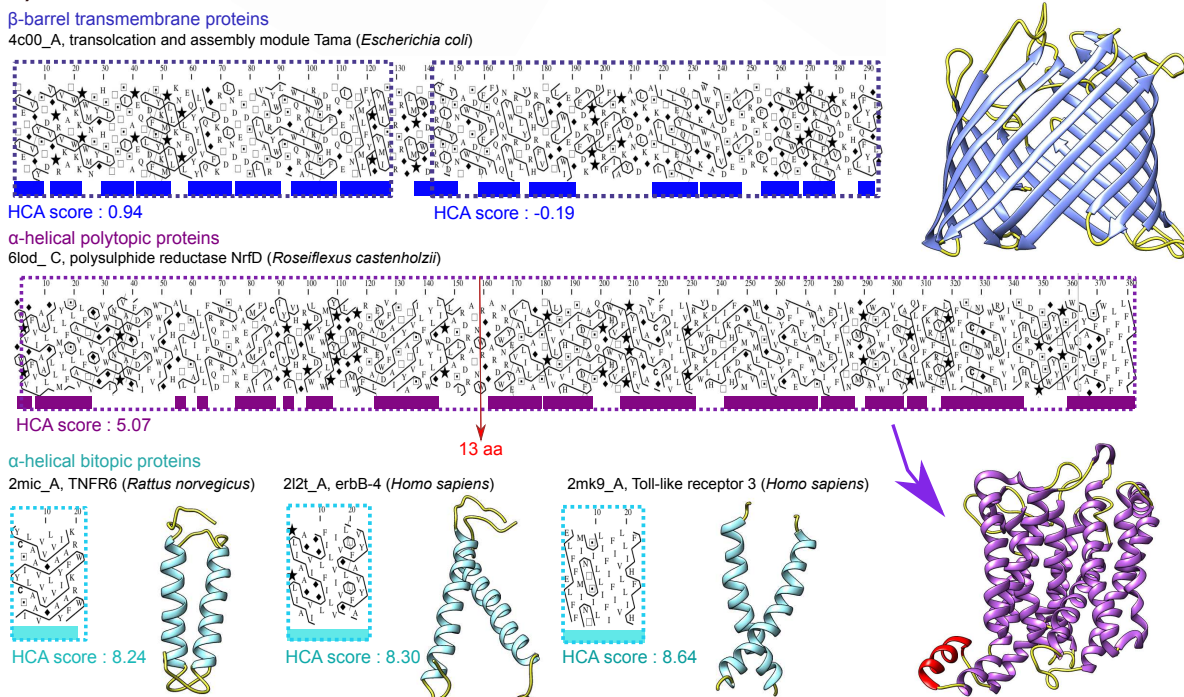
**Figure 1: Distribution of the HCA scores calculated for the foldable segments from disordered protein regions (DisProt) and ordered protein domains (soluble domains from SCOPE and membrane domains from OPM).** The considered non-redundant datasets are: DisProt v8.0.2 (HCA scores ranging from -6.95 to 9 and peaking at 0.5 and 8.1), SCOPE (HCA scores ranging from -4.67 to 9 and peaking at 1.4 and 8.1), OPM alpha-helical polytopic and bitopic categories (HCA scores ranging from -3.32 and 0.97 to 9 and peaking at 4.9 and 8.3, respectively) and OPM beta-barrel category (HCA scores ranging from -3.86 to 9 and peaking at 1.2 and 7.9). Several thresholds were fixed from these distributions, allowing to define four main classes of foldable segments (a-d). The threshold of 3.5 better discriminates the HCA scores of SCOPE foldable segments from those of alpha-helical OPM foldable segments. -1 is the threshold above which are found 95% of the HCA scores computed for foldable segments from SCOPE protein domains.

**(a)** Below -4.7: only foldable segments from disordered regions, **(b)** from -1 to 3.5, globular soluble segments (SCOPE) or membrane beta-barrels (OPM), **(c)** from 3.5 to 6.7, segments from alpha-helical transmembrane domains, **(d)** above 7.6, foldable segments composed of only one hydrophobic cluster. Two intermediate regions were also defined: **(a')** between -4.7 and -1, in this range 62% of segments are from SCOPE and 36% from DisProt, they represent only 5% of SCOPE and 11% of DisProt; **(c')** between 6.7 and 7.6, foldable segments with a dense composition in hydrophobic clusters and of short length (96% are shorter than 100 aa).

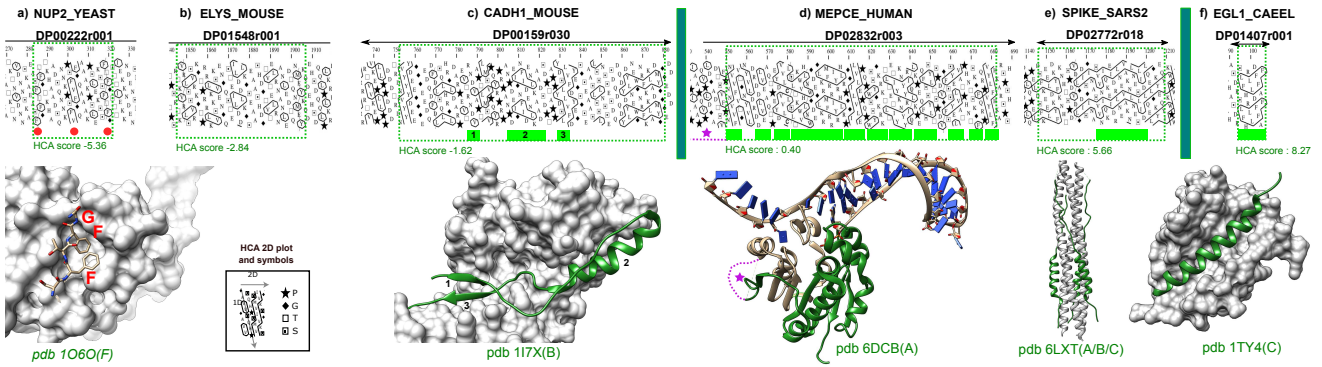
**a) SCOPe class b**



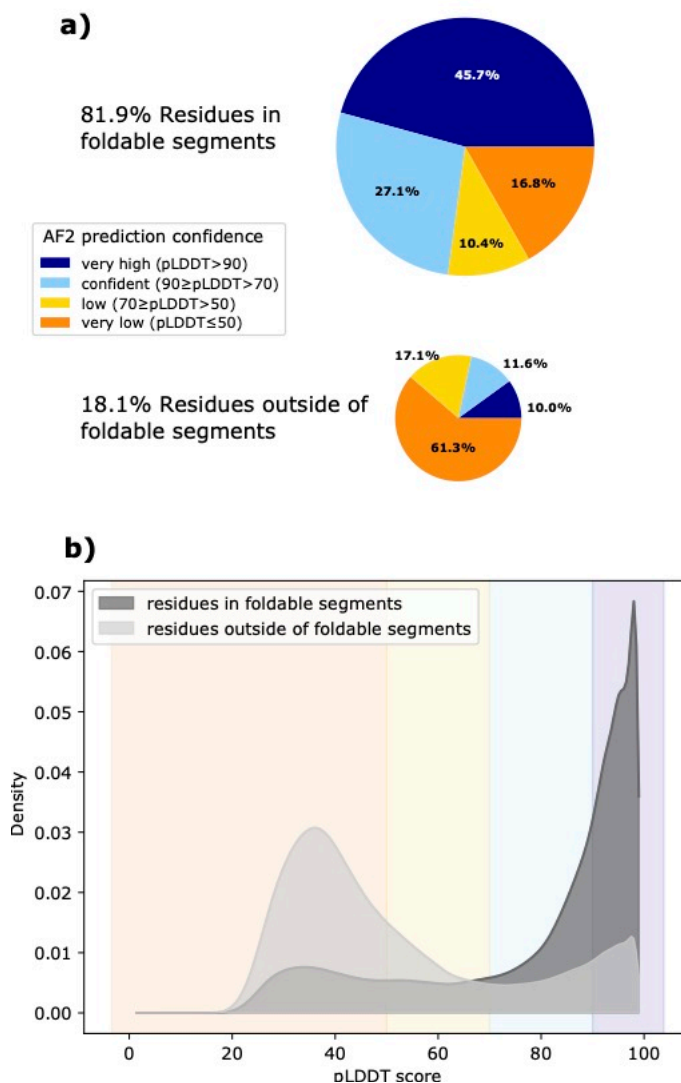
**b) OPM**



**Figure 2: Foldable segments extracted from the SCOPe and OPM databases: HCA scores, HCA 2D plots and experimental 3D structures.** The HCA plots of the sequences extracted from the SCOPe (all beta) (a) and OPM (b) databases are shown, with their foldable segments boxed (dashed lines). The corresponding HCA scores are reported below the boxes. Special symbols used in the HCA representation and the way to read the sequences and regular secondary structures (RSSs) are indicated in the inset. Positions of the RSSs, as experimentally observed from the corresponding 3D structures (ribbon representations), are reported in color below the HCA plots. The position of the 13 aa-long segment which has been removed for the OPM protein sequence in order to only keep membrane domains is highlighted in red (see *Materials and Methods* for details).

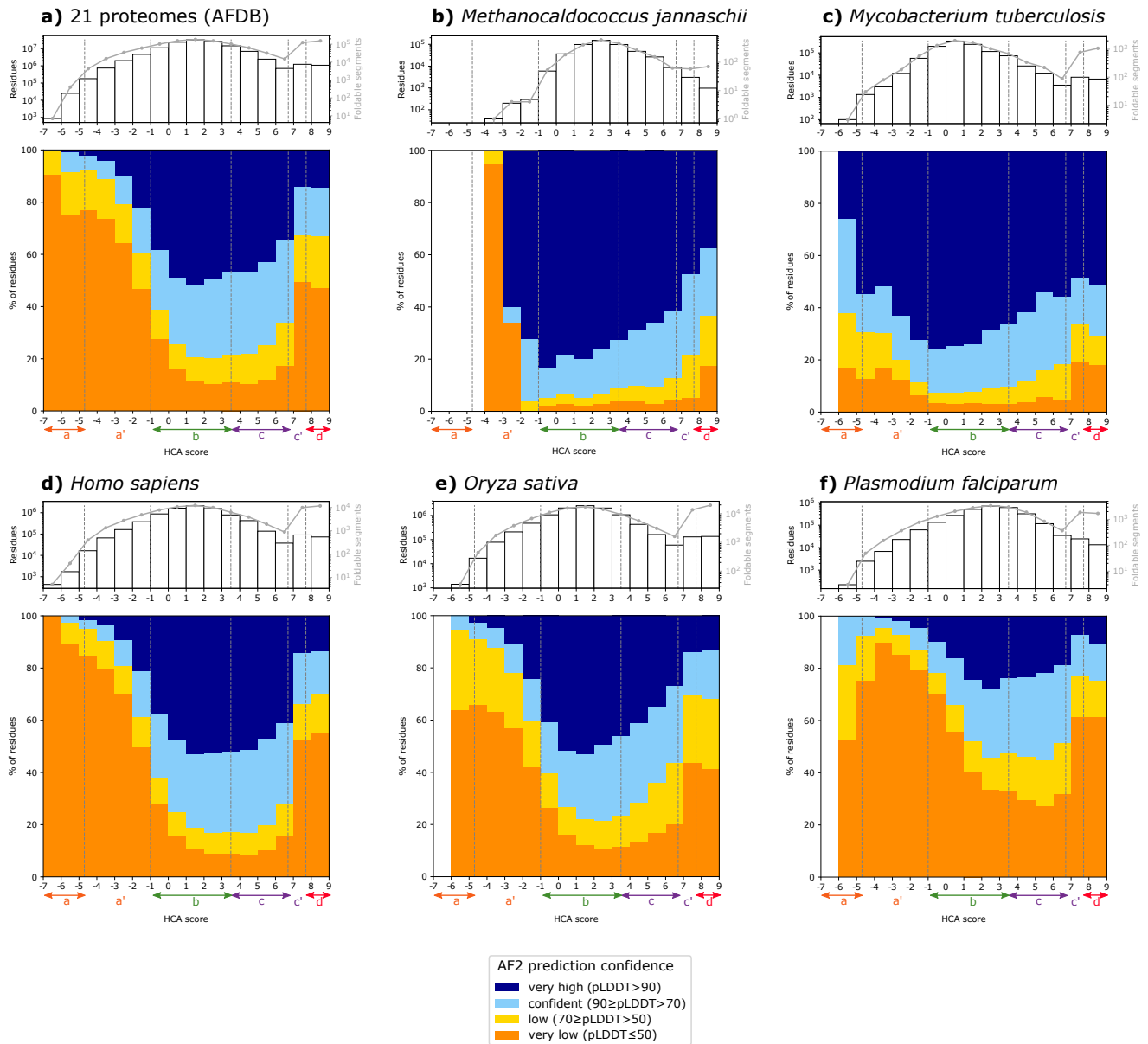


**Figure 3: Foldable segments extracted from the DisProt database: HCA scores, HCA 2D plots and experimental 3D structures.** The HCA plots of the sequences extracted from the DisProt database are shown (arrows indicate the N- or C-terminal limits of the DisProt segments). The foldable segments are boxed (dashed lines). The corresponding HCA scores are reported below the boxes. Special symbols used in the HCA representation and the way to read the sequences and regular secondary structures (RSSs) are indicated in the inset. 3D structures (in green, with their pdb identifiers) have been solved for only a few of the DisProt sequences (moreover often limited to only a part of the regions), stabilized by interaction with a partner (grey surfaces for two of them). One of the FXFG repeats of yeast NUP2 (red circles) has been modeled based on the experimental 3D structure of such a repeat in complex with importin beta-1 (*pdb 1O6O*).



**Figure 4. Distribution of AlphaFold2 per-residue prediction confidence scores (pLDDT) within and outside of foldable segments.**

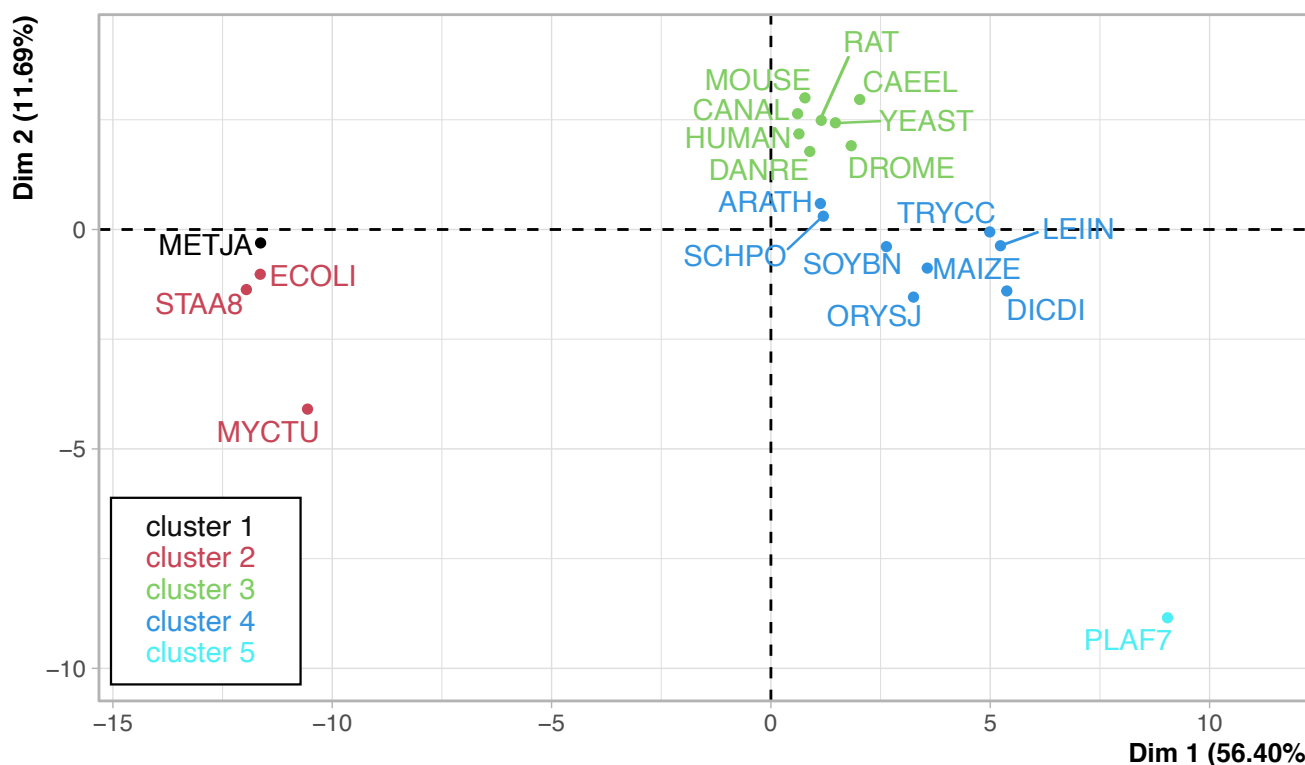
**(a)** Distribution of residues from the AFDB 21 proteomes within foldable segments (top, 81.9 % of the total) and outside of foldable segments (bottom, 18.1% of the total) in the different categories of AF2 prediction confidence. **(b)** Distribution of pLDDT scores for residues within (dark grey) and outside of (light grey) foldable segments. The AF2 prediction confidence categories are highlighted following the same color code as in (a).



**Figure 5. Relationship between HCA score and AlphaFold2 per-residue prediction confidence score for foldable segments.**

**(a)** AFDB v1 21 proteomes, **(b-f)** 5 representative proteomes from AFDB v1 (see **Figure 6** and **Supplementary Figure S6** for details).

**Top:** Representation of the number of residues by the barplot (axis on the left, base 10 logarithmic scale) and number of foldable segments by the grey line and points (axis on the right, base 10 logarithmic scale) belonging to foldable segments with HCA scores in a given interval. **Bottom:** percentage of residues in each AF2 prediction confidence categories for each HCA score range. Below the barplot are represented the foldable segment types that are likely to be found according to their HCA score, as defined in **Figure 1**: (a) disorder only, (a') mainly disorder, (b) globular domains/ membrane domains (beta-barrel), (c) membrane domains (alpha-helical), (c') mainly foldable segments with one hydrophobic cluster, (d) foldable segments with one hydrophobic cluster.

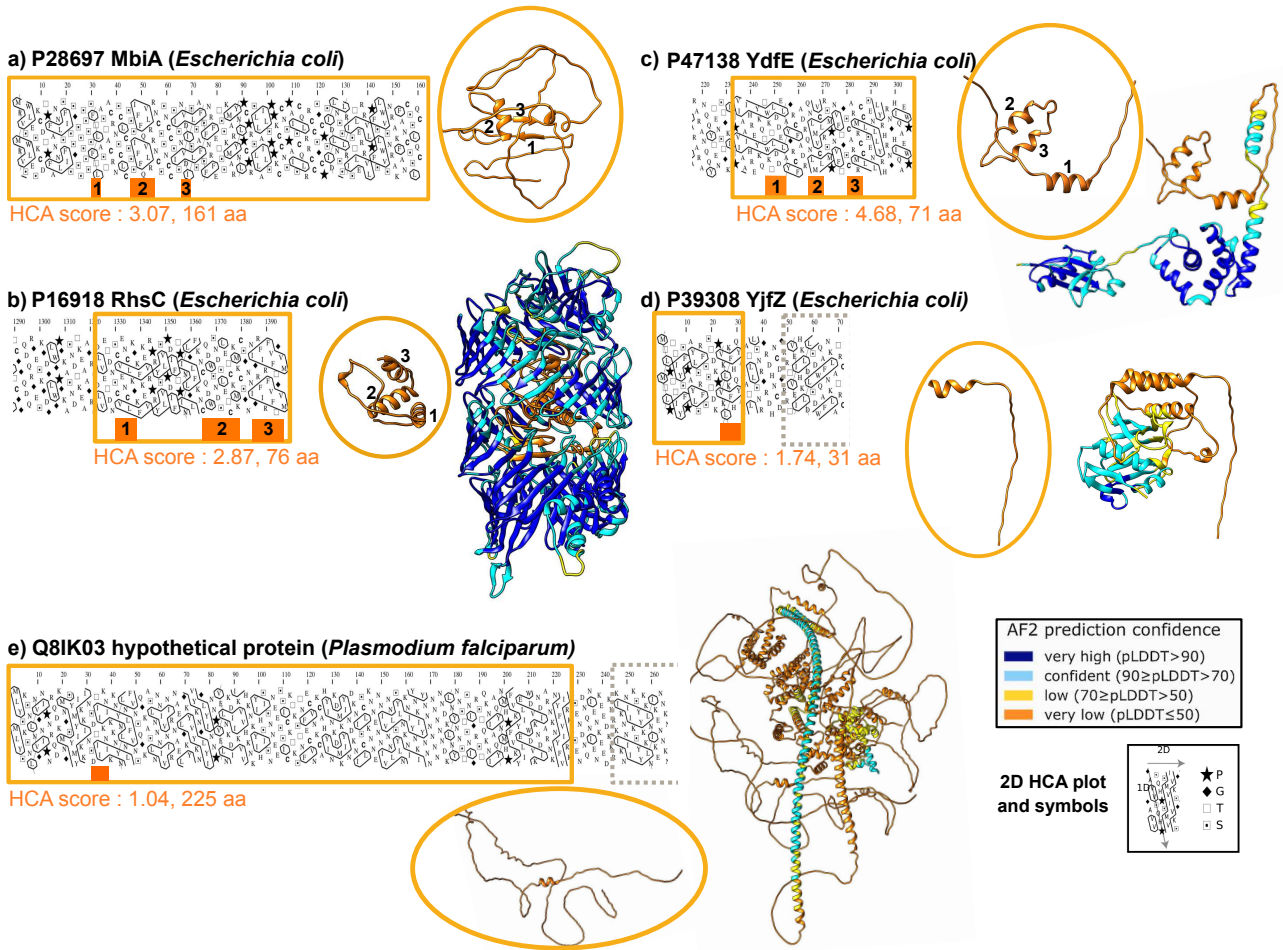


**Figure 6. Principal component analysis (PCA) for the AFDB v1 21 proteomes according to HCA score and AF2 prediction confidence.**

The 64 variables correspond to the proportion of residues of foldable segments found in each pLDDT category (very high, confident, low, very low) for each of the HCA score intervals (from -7 to 9, in steps of 1) (as represented in **Figure 4**).

The plot represents the 21 AFDB proteomes in the factor map formed by the first and second principal components (explaining 59.24% of the dataset variability). The correlation circle of the variables is represented in **Supplementary Figure S6**. Five clusters of proteomes were defined using hierarchical clustering on the first four principal components (explaining 83% of the dataset variability).

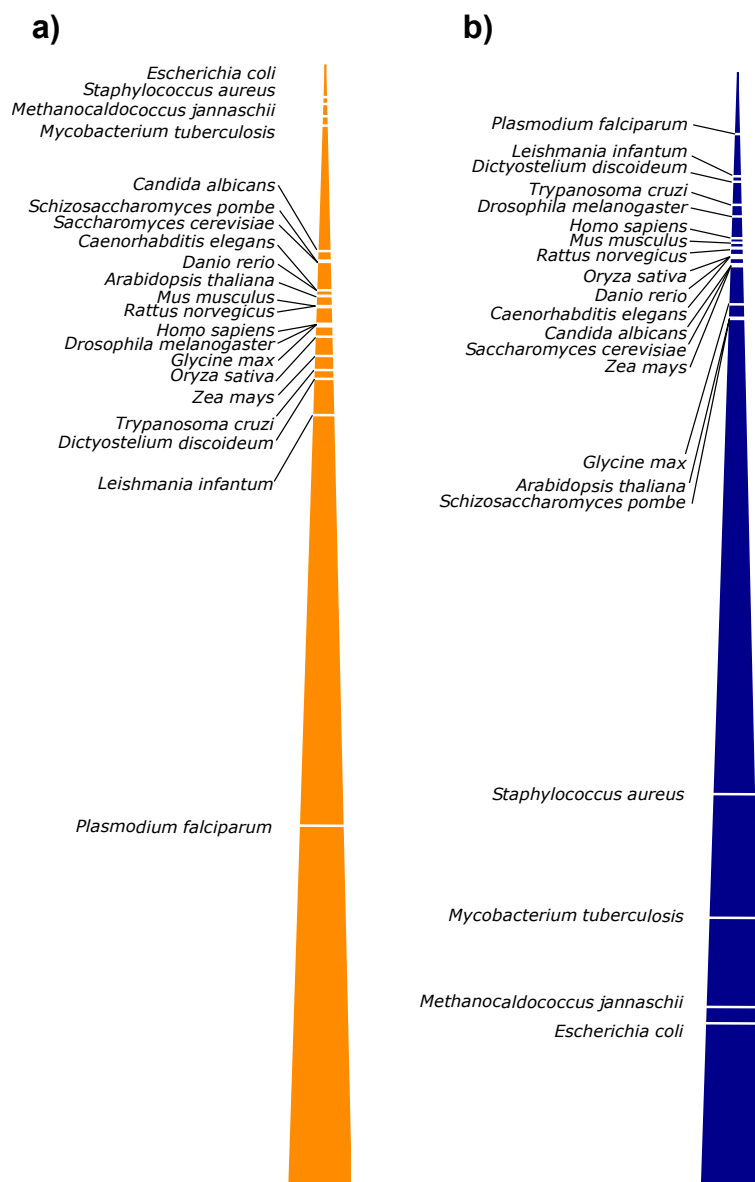
ARATH: *Arabidopsis thaliana*, CAEEL: *Caenorhabditis elegans*, CANAL: *Candida albicans*, DANRE: *Danio rerio*, DICDI: *Dictyostelium discoideum*, DROME: *Drosophila melanogaster*, ECOLI: *Escherichia coli*, HUMAN: *Homo sapiens*, LEIN: *Leishmania infantum*, MAIZE: *Zea mays*, METJA: *Methanocaldococcus jannaschii*, MOUSE: *Mus musculus*, MYCTU: *Mycobacterium tuberculosis*, ORYSJ: *Oryza sativa*, RAT: *Rattus norvegicus*, SCHPO: *Schizosaccharomyces pombe*, SOYBN: *Glycine max*, STAA8: *Staphylococcus aureus*, TRYCC: *Trypanosoma cruzi*, YEAST: *Saccharomyces cerevisiae*.



**Figure 7. Examples of very low AlphaFold2 confidence scores for globular-like domains.**

The examples shown are segments of contiguous amino acids with very low pLDDT values, fully included in foldable segments with HCA scores typical of well-folded domains (boxed in orange on the HCA plots, with the corresponding HCA scores and segment lengths (in aa) reported below). Other foldable segments of the sequences, taken from UniProt, are boxed in grey. Special symbols used in the HCA representation and the way to read the sequences and RSSs are indicated in the inset. The AF2 3D structure models of the corresponding segments are highlighted with orange circles, extracted from the models of the whole proteins (ribbon representations), colored according to the AF2 per-residue confidence metric. Positions of the RSSs, as predicted by AF2, are reported in color on the HCA plots.





**Figure 8. Classification of the AFDB v1 21 proteomes according to probable overestimation of disorder by AF2 and pyHCA in foldable and non-foldable segments, respectively.**

a) Proteomes are sorted in a top-down scale (ranging from 0 to 60%) according to the increasing proportion of residues predicted with a very low confidence by AF2 in segments with HCA scores typical of soluble globular HCA scores (category *b* defined in **Figure 1**). b) Proteomes are sorted in a top-down scale (ranging from 0 to 60%) according to the increasing proportion of residues found in non-foldable segments that are predicted with a very high confidence by AF2. For details on the proportion of residues in non-foldable segments, see **Supplementary Table S1** and **Supplementary Figure S3**.

The proteome of the archaeon *Methanocaldococcus jannaschii* harbors the lowest number of long regions (length > 30 aa) composed only by residues predicted with a very high confidence by AF2, included in non-foldable segments. The UniProt sequence accession numbers and the boundaries of these 8 regions are as follows: Q60356 [171-207], Q57673 [15-83], Q58130 [243-276], Q58560 [108-139], Q58991 [253-283], Q60317 [32-64], Q57676 [24-54] and Q58814 [28-61]. All have at least one homologous sequence with known 3D structure (data not shown).

## SUPPLEMENTARY INFORMATION

### **A sequence-based foldability score combined with AlphaFold2 predictions to disentangle the protein order/disorder continuum**

Apolline Bruley, Tristan Bitard-Feildel, Isabelle Callebaut\*, Elodie Duprat\*

*Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, 75005 Paris, France*

\* These authors should be considered as co-last authors

#### **This PDF file includes:**

- Legends for Datasets S1 to S4
- Table S1
- Dataset S1
- Figures S1 to S9
- SI References

#### **Other supplementary materials for this manuscript include the following:**

- Datasets S2 to S4

**Dataset S1.**

- A. HCA methodology, HCA plot and SEG-HCA
- B. pyHCA new functionalities

**Dataset S2 (Separate File).**

**Description of the sequences from each non-redundant dataset obtained from the reference databases SCOPe v2.0.7, OPM and DisProt v8.0.2.** Sequences are described by their length (in amino acids), HCA score, composition in strong hydrophobic amino acids (V, I, L, M, F, Y, W), sequence coverage by foldable segment, number of removed amino acids (only for OPM sequences, see Methods).

**Dataset S3 (Separate File).**

**Description of the foldable segments in each non-redundant dataset obtained from the reference databases SCOPe v2.0.7, OPM and DisProt v8.0.2.** Foldable segments are described by their length (in amino acids), HCA score, composition in strong hydrophobic amino acids (V, I, L, M, F, Y, W), number of hydrophobic clusters and associated binary code.

**Dataset S4 (Separate File).**

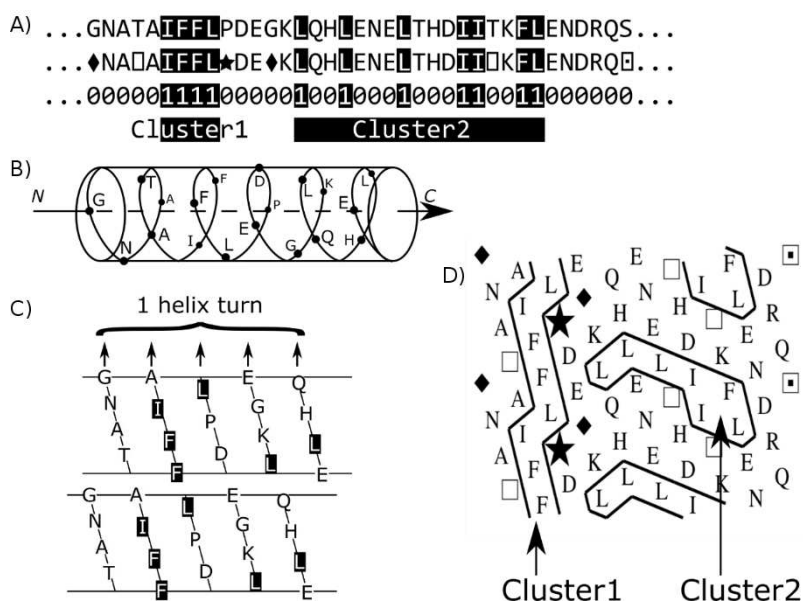
**Description of the contiguous regions of same prediction confidence from AFDB v1 sequences.** Contiguous regions are described by their length (in amino acids), the length of the foldable or non-foldable segment in which they are included, the prediction confidence category (very high, confident, low, very low), HCA score, % of charged aa.

## Dataset S1

### A. HCA methodology, HCA plot and *SEG-HCA*

HCA-based hydrophobic clusters, made of strong hydrophobic amino acids (V, I, L, F, M, Y, W), are different from simple hydrophobic motifs as they can incorporate other, non-hydrophobic residues. They are also different from simple binary patterns, being constrained by a connectivity distance defined as the minimum distance in non-hydrophobic amino acids necessary to split a binary pattern into two distinct hydrophobic clusters<sup>2</sup>. This property originates from the use of a two-dimensional alpha-helical net, connecting hydrophobic amino acids separated by up to three non-hydrophobic amino acids (or a proline)<sup>3,4</sup>. Hydrophobic clusters defined in this way (with this hydrophobic alphabet and the connectivity distance associated with the  $\alpha$ -helix) have been shown to match at best regular secondary structures ( $\alpha$ -helices and  $\beta$ -strands) and to constitute hallmarks of folded domains<sup>1,5,6</sup>.

Figure A indicates how are generated the HCA plots taken for illustration in this study. From the 1D amino acid sequence (panel A), a 2D plot is drawn (panel D) by rolling the amino acid sequence around an  $\alpha$ -helix (panel B) and cutting the helix along the horizontal axis. The helix forms a plane (panel C) on which every line of amino acids corresponds to an helix turn. The plane is duplicated and the hydrophobic clusters are defined by joining contiguous strong hydrophobic amino acids (V, I, L, F, M, Y, W)<sup>3</sup>. Regular secondary structure (RSS) elements can be easily visualized on the 2D plot, mainly corresponding to hydrophobic clusters, which are separated from each other by loops. Vertical hydrophobic clusters mainly correspond to  $\beta$ -strands and horizontal clusters to  $\alpha$ -helices. A dictionary of the most current hydrophobic clusters, established from a comprehensive analysis of experimental 3D structures, can be considered for evaluating the main propensities of hydrophobic clusters towards RSS<sup>1,5</sup>.



**Figure A.** Principle of the HCA plot. Panel A, the protein sequence (1D), in which hydrophobic amino acids are represented as white letters, is written on an  $\alpha$ -helix, displayed on a cylinder (panel B). This one is cut along the horizontal axis, unrolled and duplicated, in order to get the full environment of each amino acid, as it exists on the 1D sequence (panel C). Strong hydrophobic amino acids (V, I, L, F, M, Y, W) are encircled and their contours are joined (panel D), forming clusters. Horizontal and vertical clusters are mainly associated with alpha helices and beta strands, respectively.

Sequence segments delineated by *SEG-HCA* correspond to regions where a high density in hydrophobic clusters is detected. Hydrophobic clusters are defined here by including cysteine (C) in the commonly used hydrophobic alphabet (V, I, L, F, M, Y, W)<sup>7</sup>. *SEG-HCA* segments, referred to as foldable regions, have been shown to correspond to domains that have the ability to fold, either in an autonomous way or following contact with partners<sup>7,8</sup>. The advantage of *SEG-HCA* for the characterization of the dark

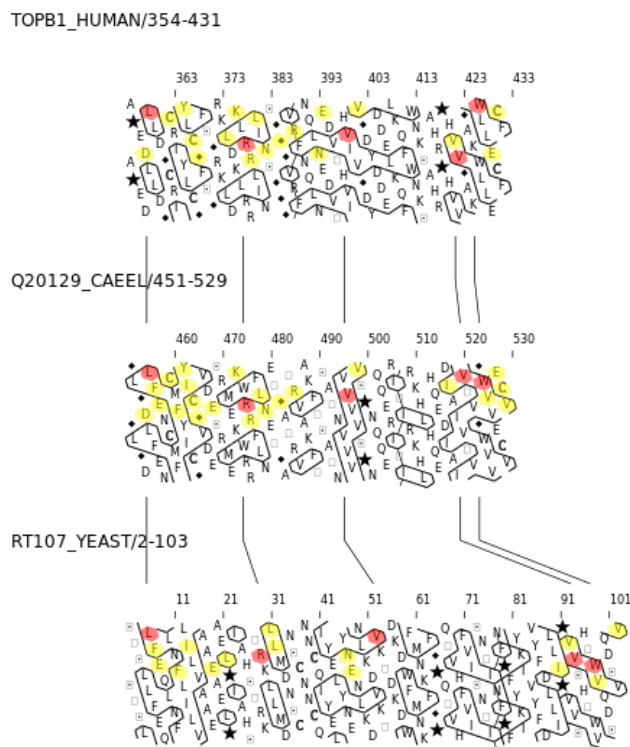
proteome is to allow the prediction of these foldable regions from the only information of a single amino acid sequence, without the prior knowledge of homologous sequences. Guidelines to interpret the HCA plots relative to the identification of foldable/disordered regions can be found in <sup>9</sup>.

## B. pyHCA new functionalities

In this section, we describe the new functionalities associated with the code refactoring of *SEG-HCA* <sup>7</sup> and *TREMOLO-HCA* <sup>10</sup> inside the pyHCA package.

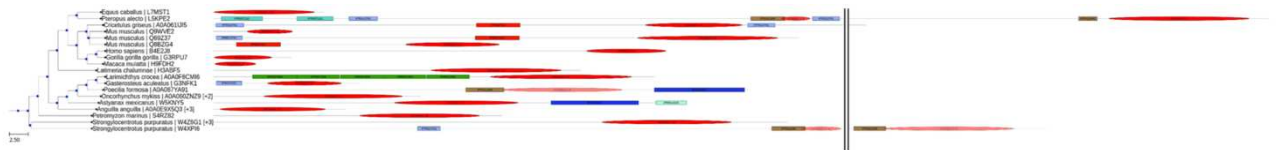
The *TREMOLO-HCA* software uses as queries regions delineated using *SEG-HCA* to search for remote similarity against protein sequences from the UniProt database <sup>11</sup> using HHblits <sup>12</sup>. For each hit, domain arrangement of the UniProt targets is retrieved from the InterPro database <sup>13</sup>. The final output allows to directly link unknown protein foldable regions, delineated by *SEG-HCA*, to existing annotations and to analyze them in the context of their domain arrangement. The original tool was based on PSI-BLAST <sup>14</sup> and the Conserved Domain Database (CDD) <sup>15</sup>. The new implementations based on HHblits and InterPro allows a more sensitive detection of protein sequence remote similarity combined with a larger coverage of the protein domain universe thanks to the multiple sources of annotations integrated in InterPro. Several scripts are also provided to easily parse and query the *TREMOLO-HCA* output and to quickly retrieve protein domains of InterPro overlapping the unknown *SEG-HCA* domains or to retrieve the whole protein domain architectures associated with the *SEG-HCA* domains.

Two drawing functionalities were also developed. The first functionality allows visualization of hydrophobic clusters of protein sequences, whether or not these are aligned (Figure B1). For each protein sequence provided in a fasta file, an HCA plot is drawn, allowing the quick inspection of its content in hydrophobic clusters, which gives information about its composition in regular secondary structures. Moreover, another new functionality was implemented to highlight conservation between aligned protein sequences on their HCA plots. Conserved protein sequence positions can therefore be inspected in the context of their hydrophobic cluster organization (Figure B1).



**Figure B1.** Example of a comparison of HCA plots of aligned protein sequences. Three members of the BRCT family, extracted from the PFAM alignment of family PF00533, are displayed. The third protein corresponds to a highly divergent member of the family, whose 3D structure has been solved (PDB: 6JOY). Red and yellow circles correspond to conserved columns in the multiple sequence alignment.

The second drawing functionality, named *domOnTree*, is built on the *TREMOLO-HCA* results and allows to easily visualize the known protein domain annotation (from InterPro) and the newly delineated domains in an evolutionary context by using the NCBI taxonomic database. Using results of *TREMOLO-HCA*, domainOnTree automatically retrieves the species associated with each hit and builds a taxonomic tree thanks to the ETE3<sup>16</sup> python package. Each leaf of the tree is then associated with its annotated sequence using InterPro annotations. Red rounded domains correspond to the *SEG-HCA* domains, which was used as query in the *TREMOLO-HCA* search, and InterPro domains are represented by rectangles with specific colors. The *SEG-HCA* domains of *TREMOLO-HCA* can then be analyzed in the context of their protein domain arrangement and visualized in terms of taxonomic specificity and domain association. An example of *domOnTree* is given in Figure B2.



**Figure B2:** DomainOnTree visualization. Using the results of *TREMOLO-HCA*, domainOnTree automatically retrieves the species associated with each hit and builds a taxonomic tree. Each leaf of the tree is then associated with its annotated sequence, using the InterPro information. Red rounded domains correspond to the *SEG-HCA* query domains, and each InterPro domain is represented by rectangle with a specific color.

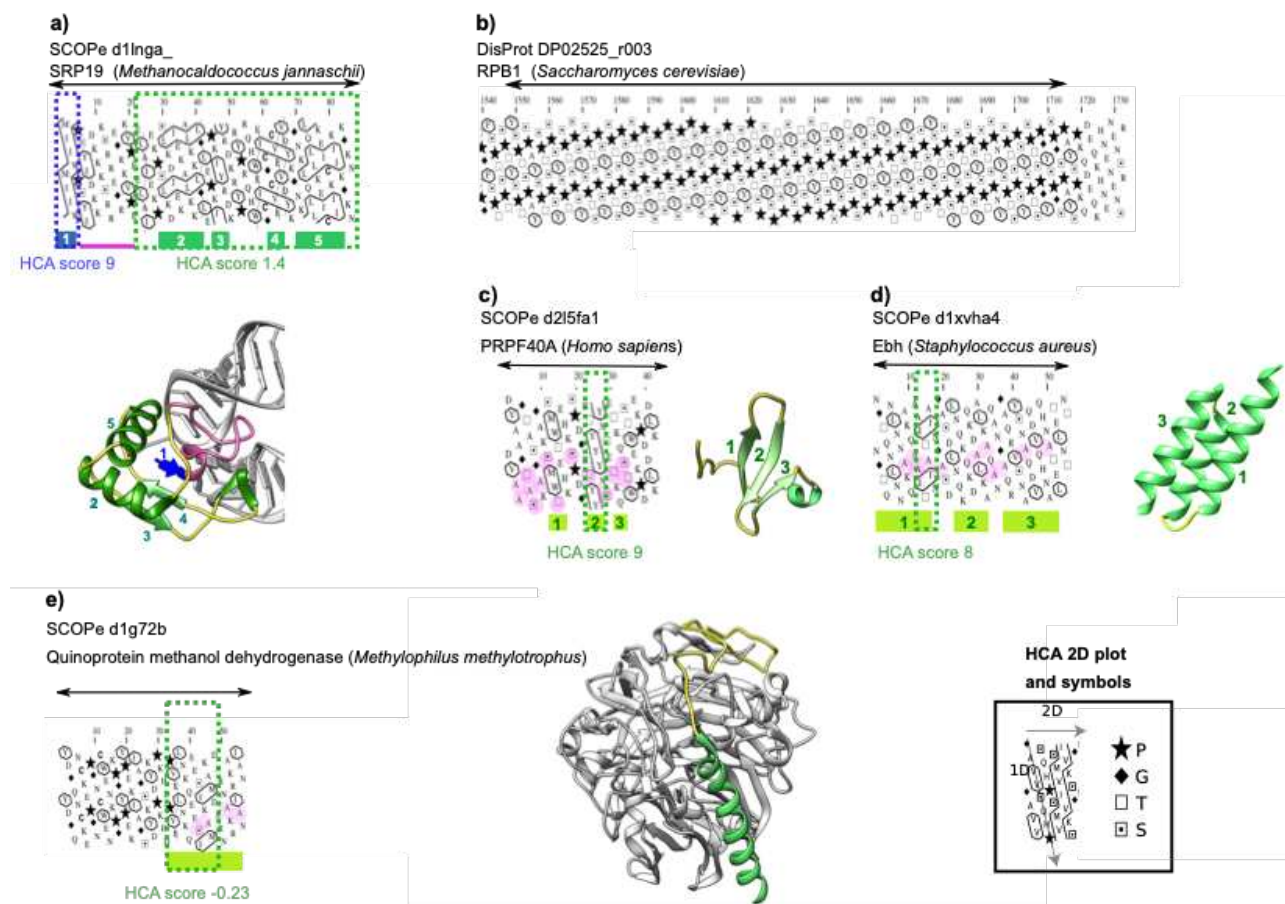
Table S1

Species	Species id in AFDB v1	Proteome id In UniProt	Sequence number	Residue number	Foldable segments					Non foldable segments				
					%aa	%VH	%C	%L	%VL	%aa	%VH	%C	%L	%VL
<i>Methanocaldococcus jannaschii</i>	METJA	UP000000805	1773	497 291	96.3	74.5	17.9	4.6	3.1	3.7	50.1	21.8	13.1	15.0
<i>Escherichia coli</i>	ECOLI	UP000000625	4363	1 349 433	91.6	74.5	19.5	4.1	1.9	8.4	51.1	21.4	13.3	14.2
<i>Staphylococcus aureus</i>	STAA8	UP000008816	2888	787 862	91.9	73.3	19.8	4.3	2.6	8.1	38.7	21.6	13.2	26.5
<i>Mycobacterium tuberculosis</i>	MYCTU	UP000001584	3988	1 313 964	83.6	72.2	18.9	5.0	3.9	16.4	45.4	18.9	13.5	22.2
<i>Arabidopsis thaliana</i>	ARATH	UP000006548	27434	11 037 703	85.4	50.4	25.4	9.8	14.4	14.6	13.2	12.7	18.9	55.2
<i>Candida albicans</i>	CANAL	UP000000559	5974	2 918 918	82.7	49.9	28.6	9.0	12.5	17.3	10.0	12.1	14.6	63.3
<i>Schizosaccharomyces pombe</i>	SCHPO	UP000002485	5128	2 345 773	86.4	49.0	28.9	9.2	12.9	13.6	13.2	14.5	16.3	56.0
<i>Saccharomyces cerevisiae</i>	YEAST	UP000002311	6040	2 902 659	84.0	47.8	29.4	9.3	13.4	16.0	10.3	12.9	14.5	62.4
<i>Mus musculus</i>	MOUSE	UP000000589	21613	10 879 808	80.8	47.7	27.6	8.7	16.0	19.2	9.1	11.3	13.9	65.7
<i>Rattus norvegicus</i>	RAT	UP000002494	21270	10 424 727	80.8	47.7	27.6	8.8	15.9	19.2	9.4	11.3	14.3	65.0
<i>Danio rerio</i>	DANRE	UP000000437	24661	12 654 340	82.0	47.5	27.9	8.7	15.8	18.0	9.8	11.7	13.9	64.6
<i>Glycine max</i>	SOYBN	UP000008827	55799	21 578 168	85.8	47.2	25.8	10.9	16.1	14.2	12.5	12.4	19.2	56.0
<i>Homo sapiens</i>	HUMAN	UP000005640	20294	10 532 415	79.4	46.3	27.5	8.7	17.5	20.6	8.9	11.3	14.0	65.8
<i>Caenorhabditis elegans</i>	CAEEL	UP000001940	19694	7 812 030	84.7	45.8	29.8	10.9	13.5	15.3	9.9	13.9	19.1	57.1
<i>Drosophila melanogaster</i>	DROME	UP000000803	13458	6 656 778	79.0	45.6	27.1	9.7	17.7	21.0	7.7	10.6	14.3	67.3
<i>Oryza sativa</i>	ORYSJ	UP000059680	43649	13 374 028	78.4	45.5	24.4	12.4	17.7	21.6	9.5	10.8	25.0	54.8
<i>Zea mays</i>	MAIZE	UP000007305	39299	15 340 862	80.5	42.8	27.1	11.8	18.3	19.5	10.4	11.8	20.1	57.7
<i>Dictyostelium discoideum</i>	DICDI	UP000002195	12622	6 507 765	80.2	38.4	29.5	13.1	19.0	19.8	5.9	10.5	15.0	68.6
<i>Trypanosoma cruzi</i>	TRYCC	UP000002296	19036	8 959 621	79.4	36.3	31.6	13.0	19.1	20.6	7.1	11.7	19.5	61.7
<i>Leishmania infantum</i>	LEIIN	UP000008153	7924	4 678 533	73.1	34.3	30.3	11.9	23.5	26.9	5.6	10.1	11.0	73.4
<i>Plasmodium falciparum</i>	PLAF7	UP000001450	5187	3 343 123	84.8	22.5	24.7	13.1	39.7	15.2	3.3	6.6	8.6	81.5

**Description of the 21 AFDB proteomes by their coverage in foldable segments in relation to the prediction confidence from AlphaFold2.**

3104 sequences were excluded from our analyses, as sequences longer than 2700 amino acids were not modeled in AFDB v1. The AF2 prediction confidence categories are abbreviated as follows: VH stands for very High, C for confident, L for low and VL for very low.

Figure S1



### Coverage of sequences from reference databases by foldable segments.

The limits of the sequences in the reference databases are indicated with arrows on the HCA plots of the sequences, while foldable segments are boxed with dashed lines, with the corresponding HCA scores reported below. Special symbols used in the HCA representation and the way to read the sequences (1D) and regular secondary structures (2D) are indicated in the inset. Positions of the RSSs, as experimentally observed from the corresponding 3D structures (ribbon representations), are reported in color below the HCA plots.

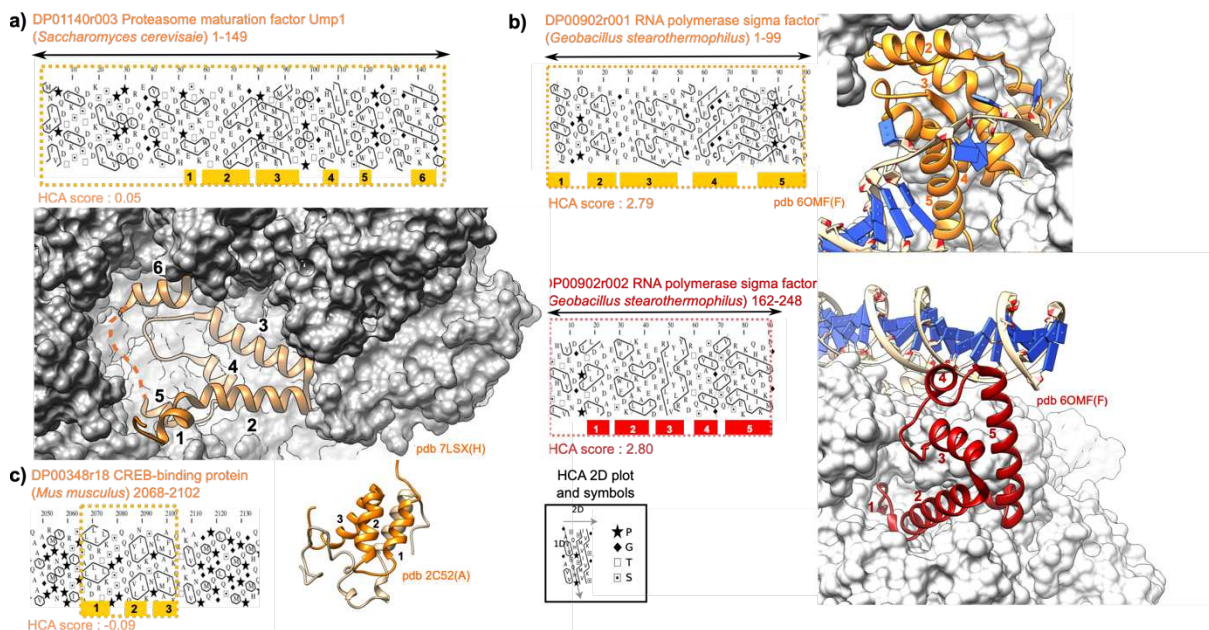
a) Example of a sequence from the SCOPe database and partially covered by foldable segments.

b) Example of a sequence from the DisProt database with no foldable segment.

c-e) Examples of sequences of the SCOPe database with a low coverage in foldable segments. The two first examples (c and d) correspond to well-folded segments, with non-strong hydrophobic amino acids (alanine, serine and threonine (pink shaded)) participating in the hydrophobic core. The third example (e) corresponds to a subunit of an enzyme, which folds in the context of the dimer in which it participates (the other subunit is shown in grey). The pdb file 1XVH has been superseded by 4KJM.



Figure S2



**Examples of DisProt sequences totally, or almost totally covered by foldable segments, with properties of soluble globular domains.**

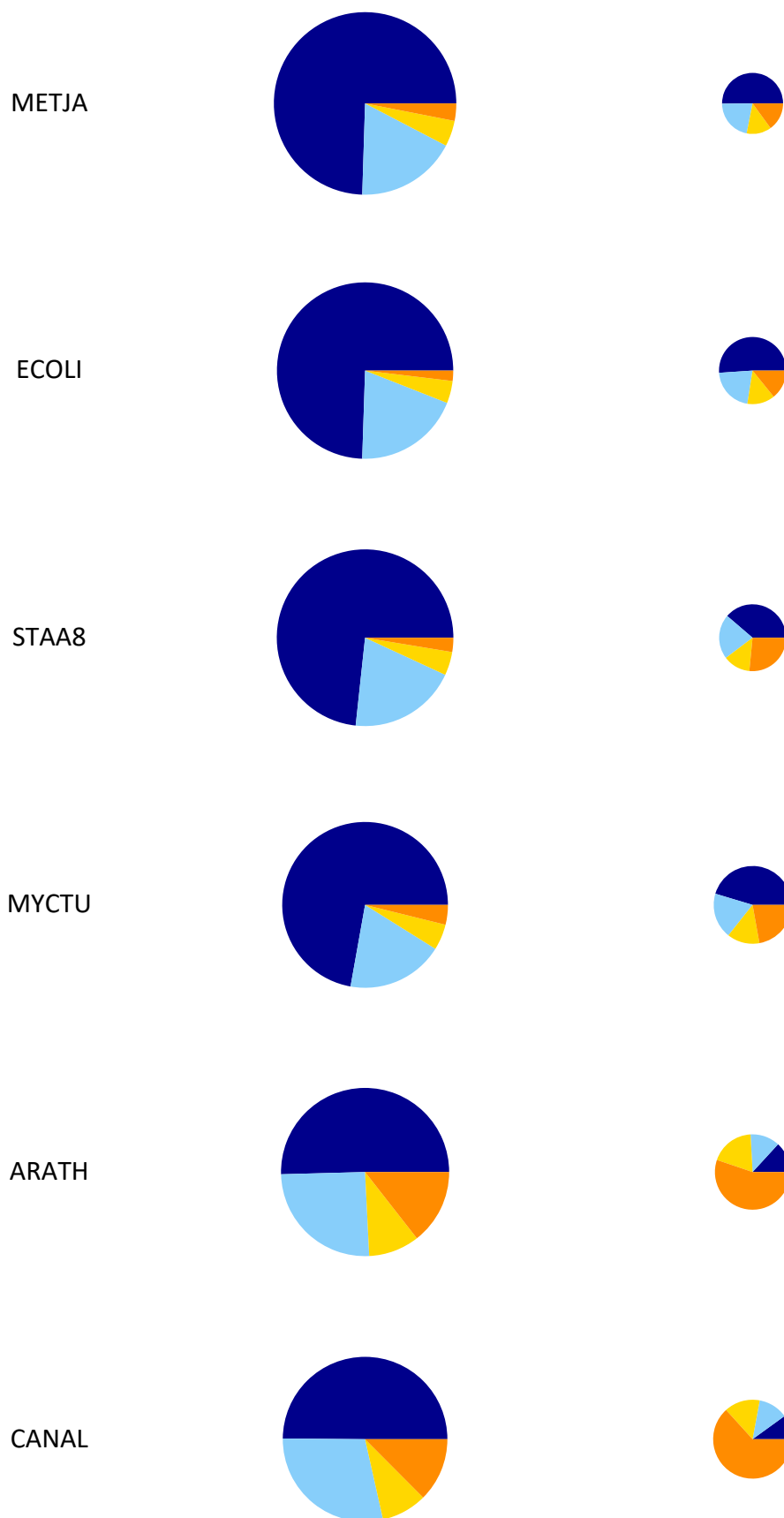
The sequences extracted from the DisProt database are designated with arrows on the HCA plot, while the foldable segments are boxed with dashed lines, with the corresponding HCA scores reported below. Special symbols used in the HCA representation and the way to read the sequences (1D) and regular secondary structures (2D) are indicated in the inset. Positions of the regular secondary structures, as experimentally observed from the corresponding 3D structure (ribbon representations), are reported in color below the HCA plots.

a) The yeast proteasome maturation factor Ump1 is disordered when free in solution, as observed using various experimental techniques<sup>17,18</sup>. It however forms well-ordered secondary structures in complex with the 20S core particle, playing a key role in the dynamical assembly of proteasome<sup>19</sup> – pdb 7LSX (chain H)).

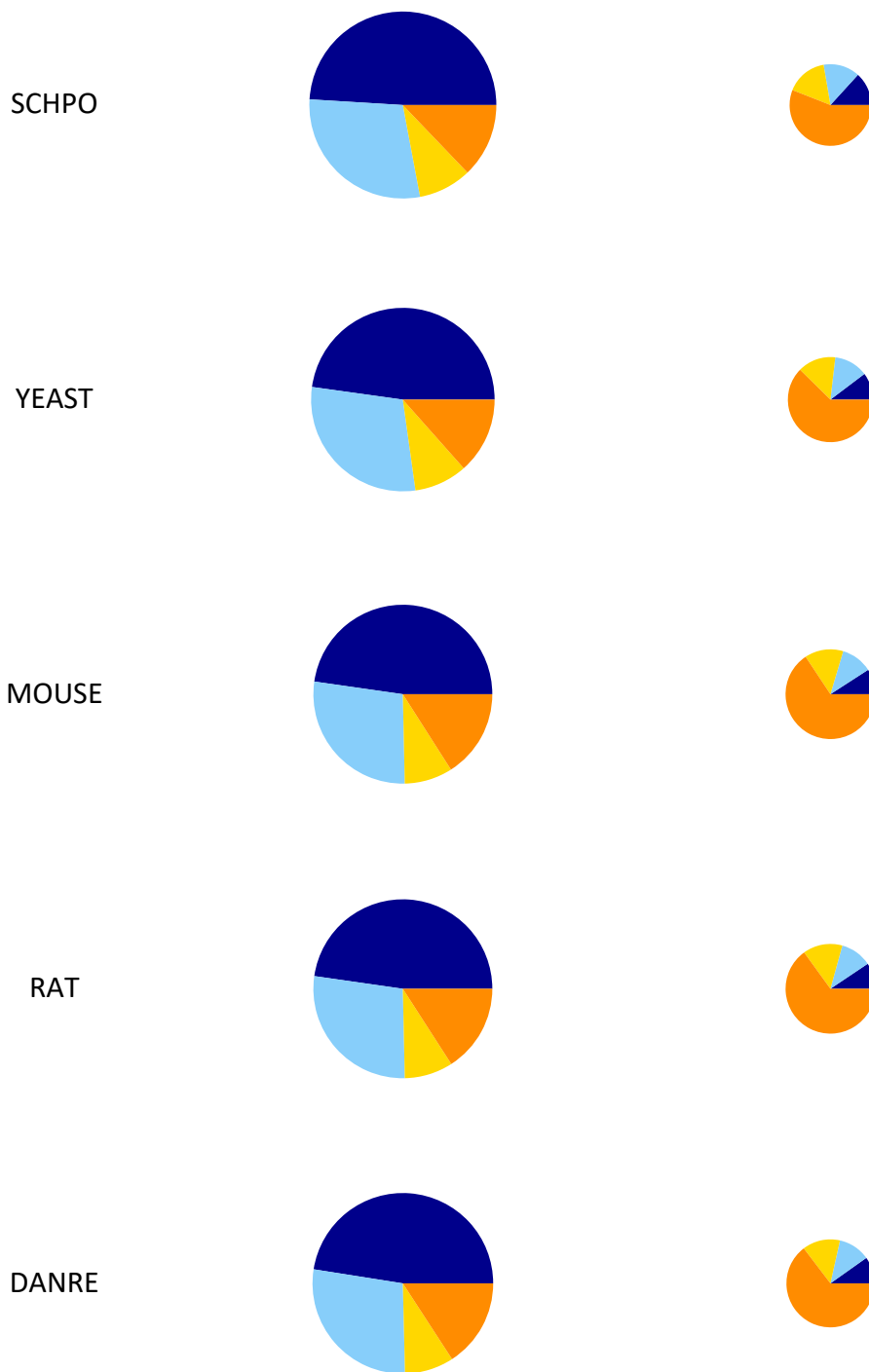
b) The *Geobacillus stearothermophilus* RNA polymerase sigma factor has been described as disordered<sup>(20)</sup> - DP00902r001 (aa 1-101) and DP00902r002 (aa 159-250)). In a further article, these two sequences were observed well folded (orange and red, respectively) in the *E. coli* SigmaS-transcription initiation complex with activator Crl<sup>(21)</sup> - pdb 6OMF (chain F)).

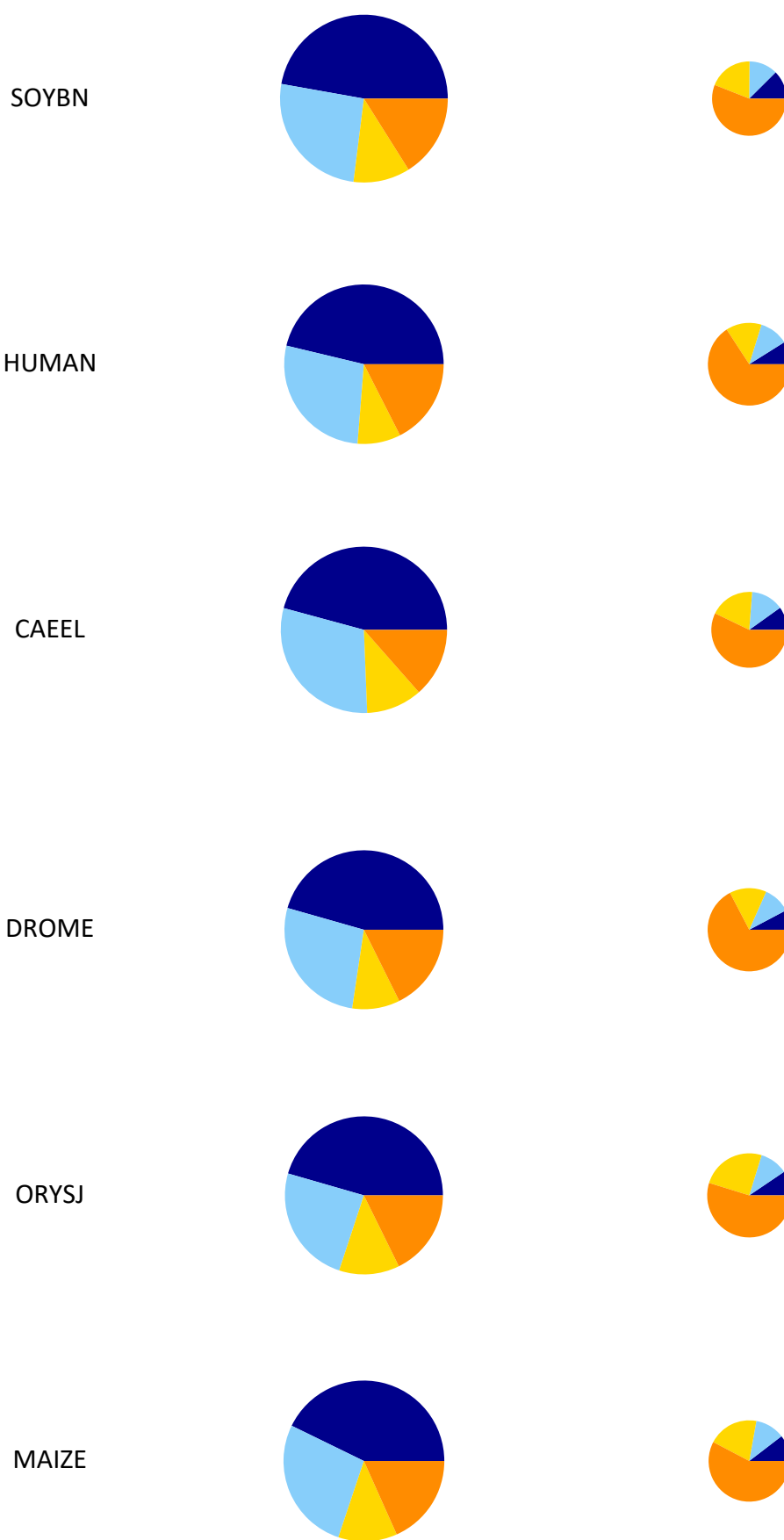
c) The mouse CREB-binding protein (DisProt DP00348r018 aa 2059-2117; foldable segment aa 2068-2102, HCA score = -0.09), in complex with human nuclear receptor coactivator 1 (NCOA1 – pdb 2C52).

**Figure S3**

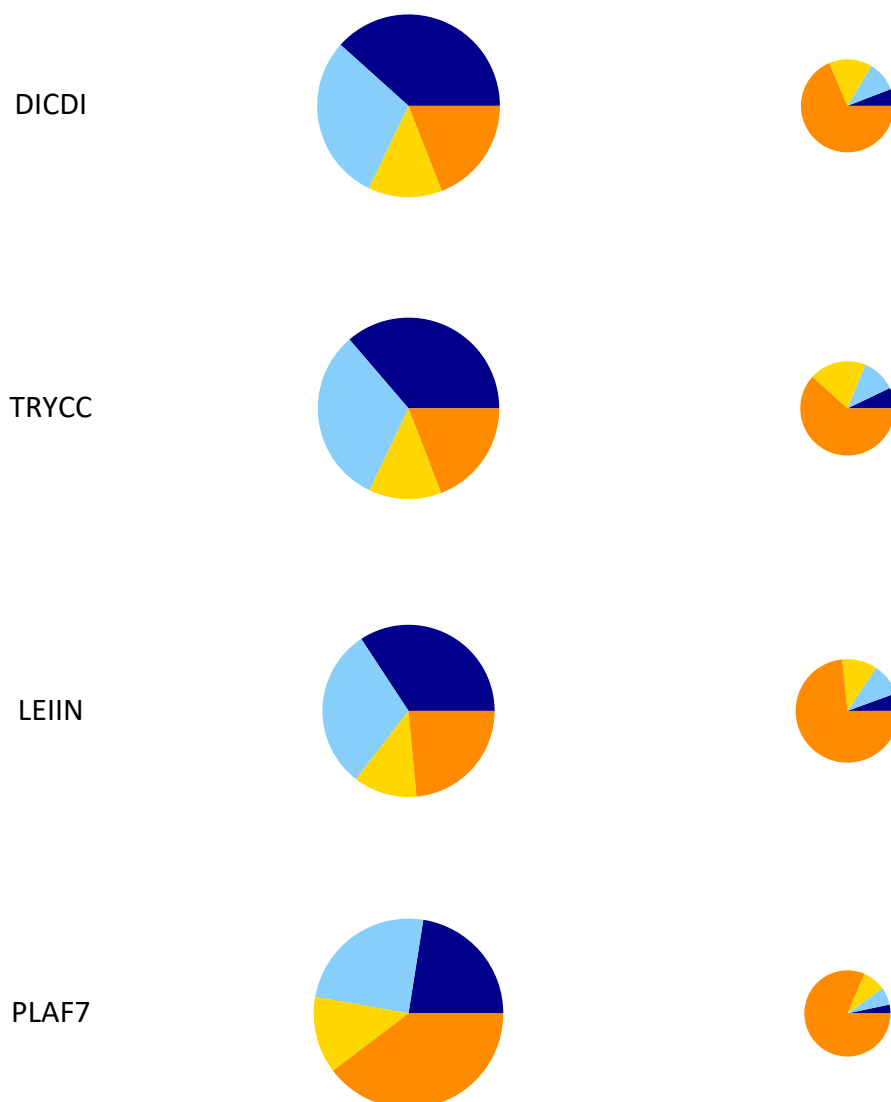


A sequence-based foldability score combined with AlphaFold2 predictions to disentangle the protein order/disorder continuum



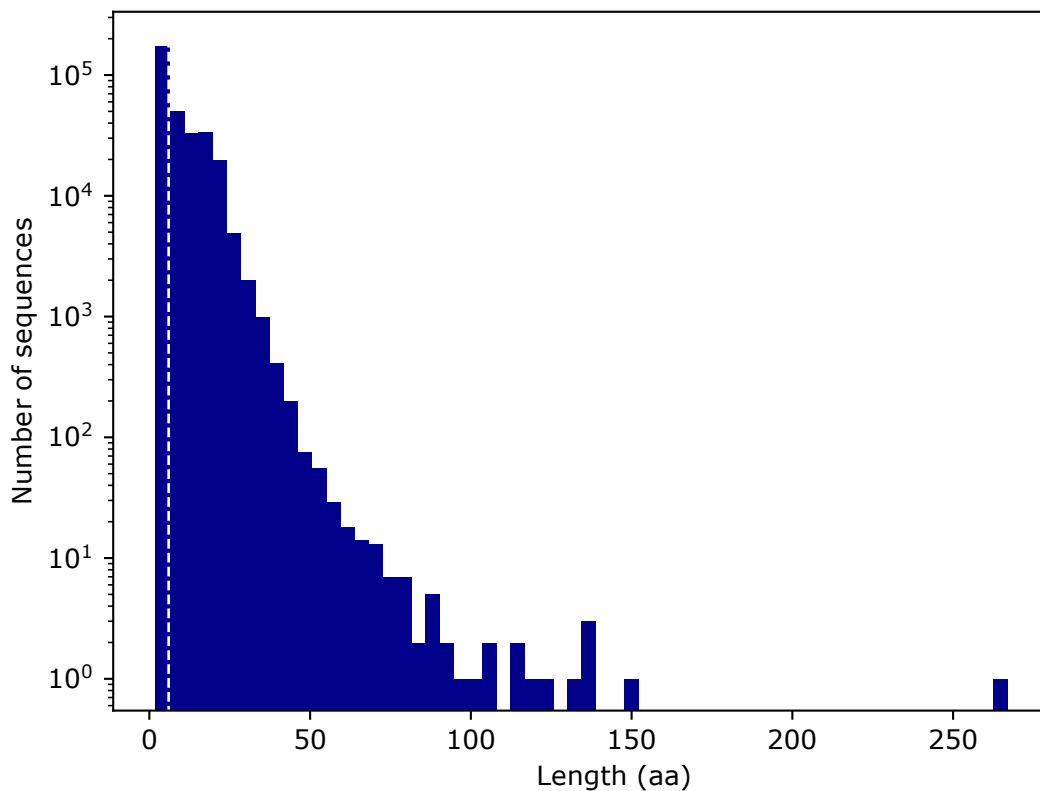


A sequence-based foldability score combined with AlphaFold2 predictions to disentangle the protein order/disorder continuum



**Distribution of residues from each of the 21 AFDB v1 proteomes within foldable segments (left) and outside of (right) foldable segments in the different categories of the AF2 pLDDT.**

Pie charts are sized according to the relative proportion of residues within foldable segments or outside of foldable segments. The proteomes are sorted by the proportion of residues belonging to foldable segments that are predicted by AF2 with very high confidence (dark blue). The values represented here and the complete organism names can be found in **Supplementary Table S1**.

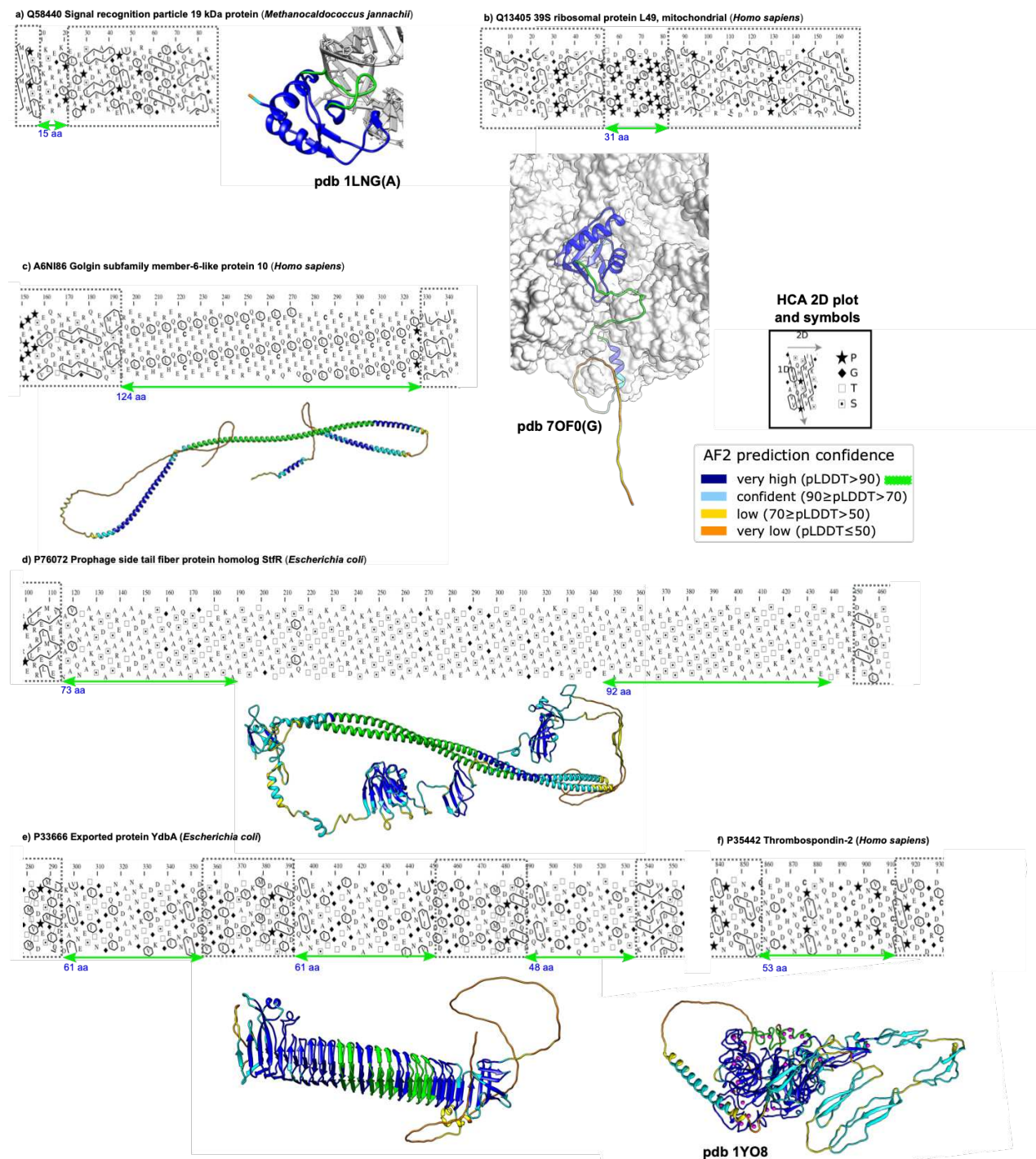
**Figure S4**

**Length of contiguous segments with very high AF2 prediction confidence located outside of foldable segments.**

Distribution of the length of contiguous segments of AF2 very high confidence prediction (2 or more contiguous residues of pLDDT>90) found outside of foldable segments (429 605 segments represented). The white dashed line represents the median length (6 amino acids). In this dataset, 13 segments are composed of more than 100 amino acids, distributed in 6 proteomes: *Homo sapiens* (5 segments), *Caenorhabditis elegans* (3 segments), *Mycobacterium tuberculosis* (2 segment), *Dictyostelium discoideum* (1 segment), *Drosophila melanogaster* (1 region), *Oryza sativa* (1 segment).

A sequence-based foldability score combined with AlphaFold2 predictions to disentangle the protein order/disorder continuum

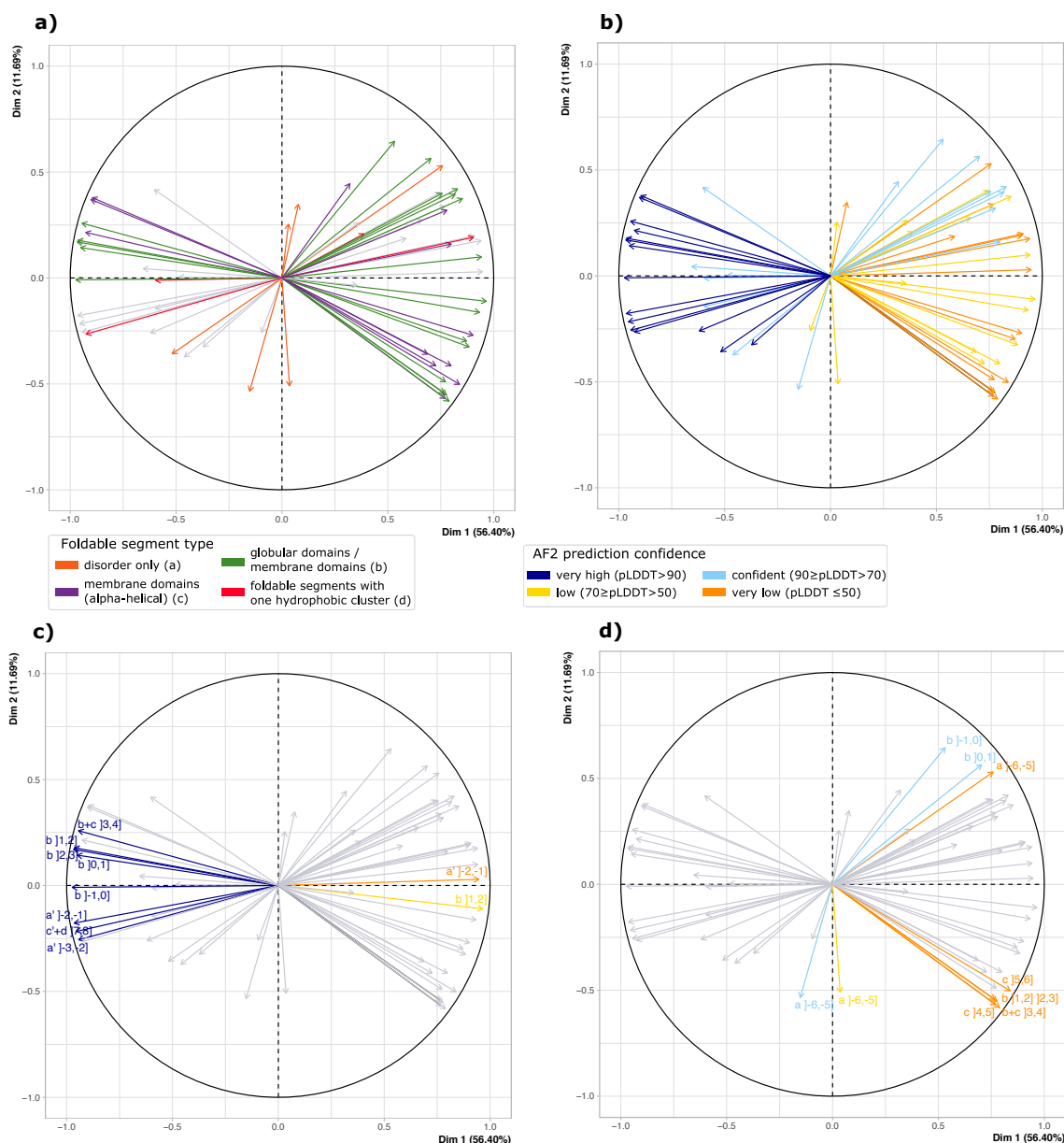
Figure S5



### Regions with high AlphaFold2 pLDDT outside of foldable segments.

Examples of segments of contiguous amino acids outside of foldable segments, with very high pLDDT scores. These segments are highlighted with green arrows on the HCA plots of the sequences taken from UniProt, with the segment lengths (in aa) reported below. Foldable segments are boxed in grey. Special symbols used in the HCA representation and the way to read the sequences and RSSs are indicated in the inset. The corresponding segments are highlighted in green in the AlphaFold2 3D structure models of the whole proteins (ribbon representations), colored according to the AlphaFold2 per-residue confidence metrics in the other regions. In panels a, b and e, the AlphaFold2 models were superimposed to the corresponding experimental 3D structures solved in complex with partners (nucleic acids, ribosomal complex and calcium ions, respectively).

Figure S6



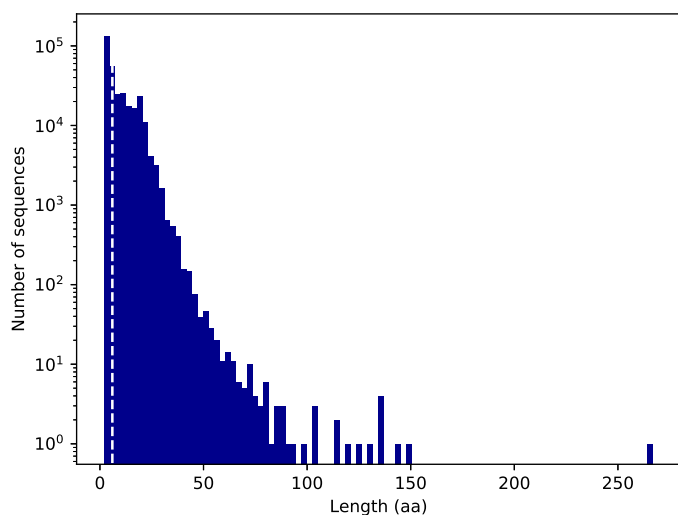
### Principal component analysis (PCA) for the AFDB v1 21 proteomes according to the HCA score and AF2 prediction confidence.

The 64 variables correspond to the proportion of residues of foldable segments found in each pLDDT category (very high, confident, low, very low) for each of the 16 HCA score intervals (from -7 to 9 with a step of 1) (as represented in **Figure 5**). Each principal component (or axis) is a linear combination of variables defined in order to maximize the discrimination of the proteomes; coordinates of vectors on each axis indicate the coefficient of the variables in the linear combination. The smaller the angle between arrows, the stronger the positive linear correlation between the corresponding variables. Orthogonal arrows indicate no correlation, while arrows pointing in opposite directions indicate a negative correlation. **(a)** All 64 variables colored according to the type of foldable segments that predominates for their HCA score (as defined in **Figure 1**), in grey are represented the categories 'a' and 'c'. **(b)** All 64 variables colored according to the AF2 prediction confidence. **(c-d)** Representation of the 10 variables with the highest contribution to the first **(c)** and second **(d)** principal component in color (same color code as **(b)**). The labels correspond to the foldable segment HCA score interval and to the type of foldable segments found in this interval, as described in **Figure 1**: a- disorder only, b- soluble globular, c- alpha transmembrane, d- only foldable segments with one hydrophobic cluster.

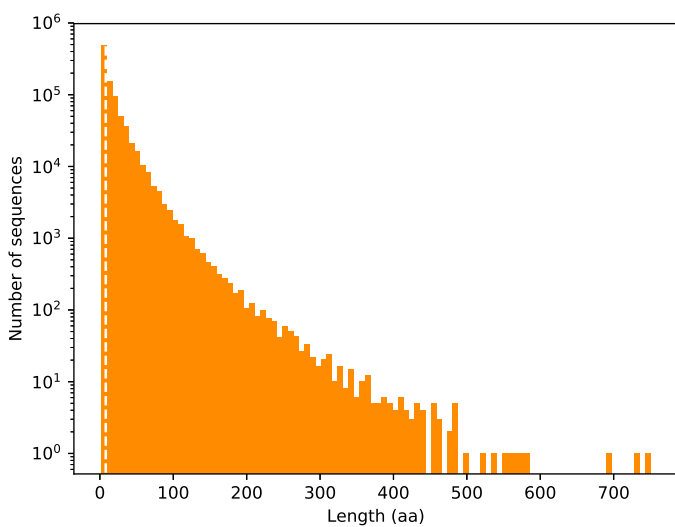


**Figure S7**

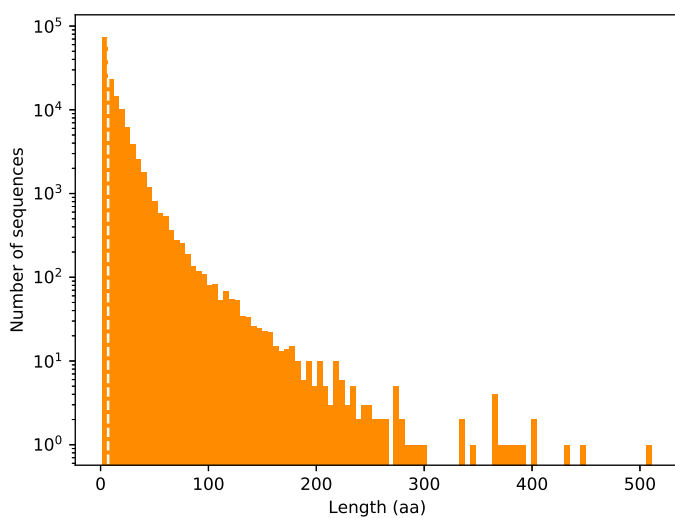
a) Disordered-like regions with very high AF2 confidence prediction



b) Soluble-like regions with very low AF2 confidence prediction



c) Transmembrane-like regions with very low AF2 confidence prediction

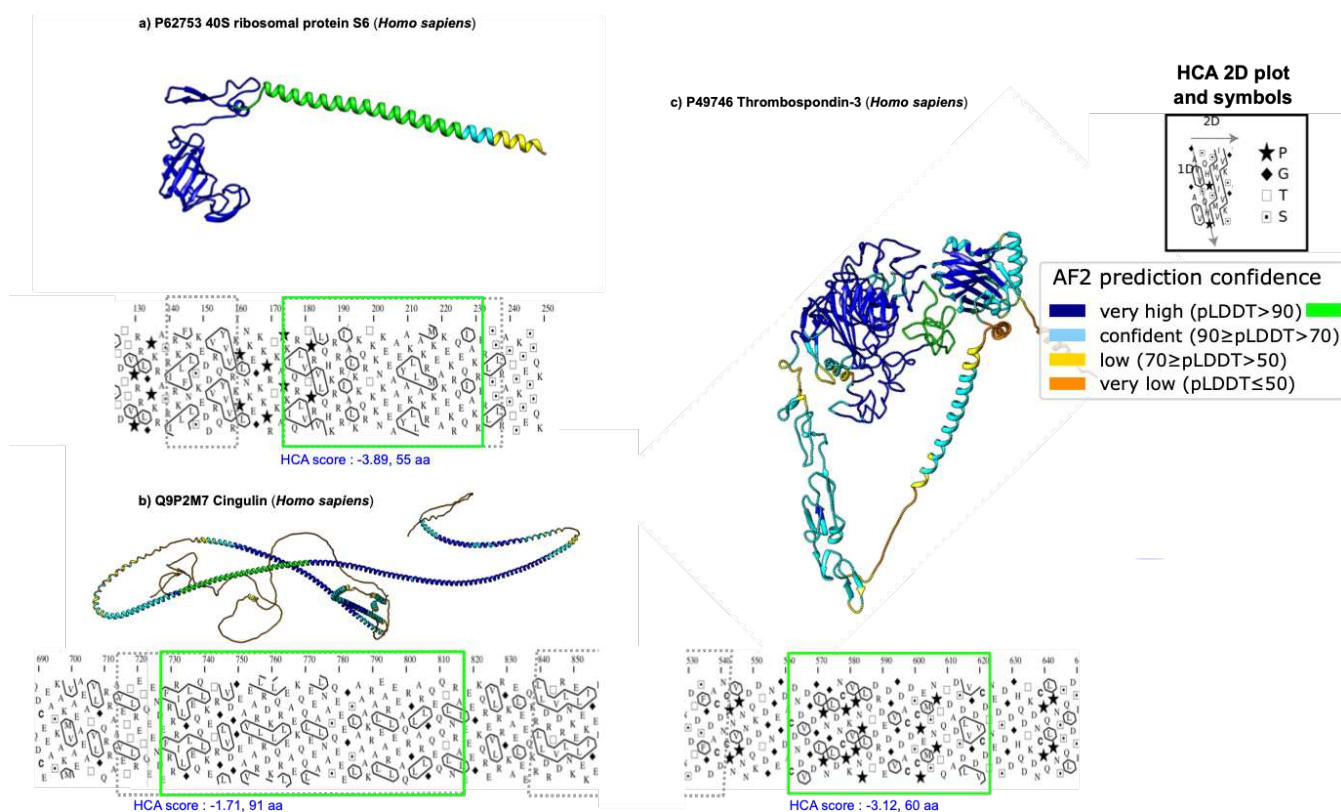


**Length of contiguous segments with similar AF2 prediction confidence located in foldable segments.**

(a) Distribution of the lengths of contiguous segments of AF2 very high confidence predictions (2 or more contiguous residues of pLDDT>90) found in foldable segments with HCA score below -4.7 (corresponding to the disorder category - 319840 segments). (b) Distribution of the lengths of contiguous segments of AF2 very low confidence predictions (2 or more contiguous residues of pLDDT≤50) found in foldable segments with HCA score between -1 and 3.5 (corresponding to the soluble globular segment category - 896049 segments). (c) Distribution of the lengths of contiguous segments of AF2 very low confidence predictions (2 or more contiguous residues of pLDDT≤50) found in foldable segments with HCA score between 3.5 and 6.7 (corresponding to the transmembrane segment category - 141742 segments). The white dashed lines represent the median lengths (8 amino acids for both distributions).

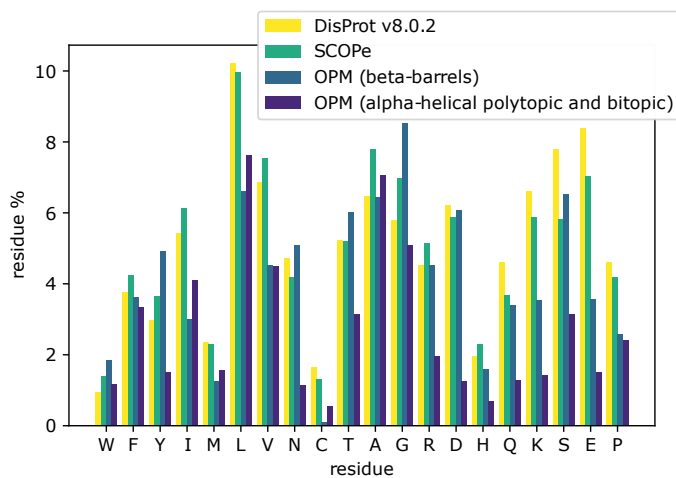
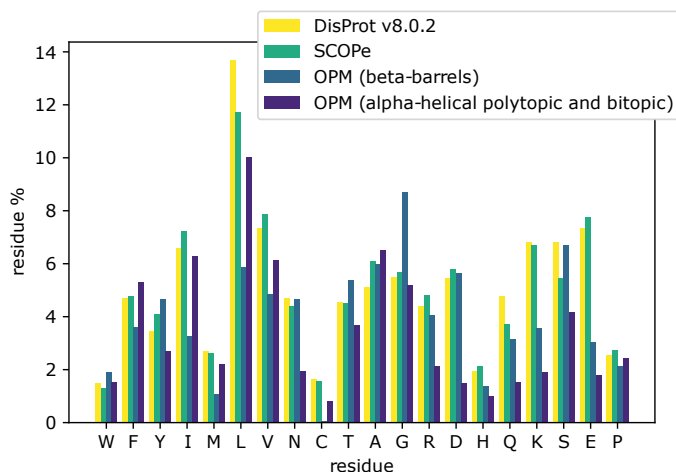
A sequence-based foldability score combined with AlphaFold2 predictions to disentangle the protein order/disorder continuum

Figure S8



**Regions with high AlphaFold2 pLDDT in foldable segments with low HCA scores.**

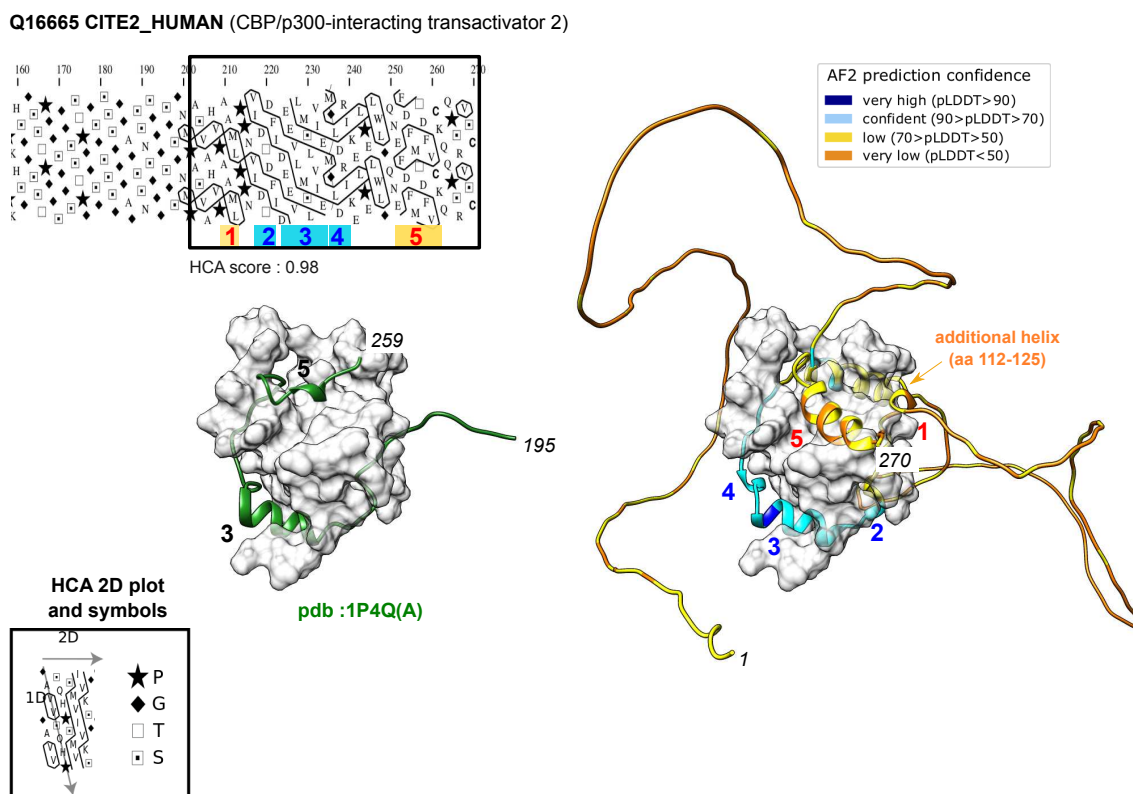
Examples of segments of contiguous amino acids within foldable segments with very high pLDDT and low HCA scores. These segments are boxed in green on the HCA plots of the sequences taken from UniProt, with the corresponding HCA scores and segment lengths (in aa) reported below. These segments cover (totally or for a very large part) foldable segments, which are boxed in grey. Special symbols used in the HCA representation and the way to read the sequences and RSSs are indicated in the inset. The corresponding segments are highlighted in green in the AlphaFold2 3D structure models of the whole proteins (ribbon representations), colored according to the AlphaFold2 per-residue confidence metrics in the other regions.

**Figure S9****a)****b)**

**Amino acid composition of foldable segments (of length > 30aa) with HCA scores typical of globular sequences (regions b and c) from disordered sequences (DisProt v8.0.2) or globular sequences (soluble domains from SCOPe and transmembrane domains from OPM).**

**(a)** Foldable segments typical of soluble, globular domains ( $-1 \leq \text{HCA score} \leq 3.5$ , structural state b as defined in **Figure 4**) from DisProt, SCOPe, OPM beta-barrels, OPM alpha-helical polytopic and bitopic segments (624, 10806, 164, 135 segments, respectively). **(b)** Foldable segments with HCA scores typical of alpha-helical transmembrane segments ( $3.5 < \text{HCA score} \leq 6.7$ , structural state c) from DisProt, SCOPe, OPM beta-barrels, OPM alpha-helical polytopic and bitopic segments (87, 1578, 32, 671 segments, respectively). The amino acids are ordered from order promoting to disorder promoting, as proposed in <sup>22</sup>.

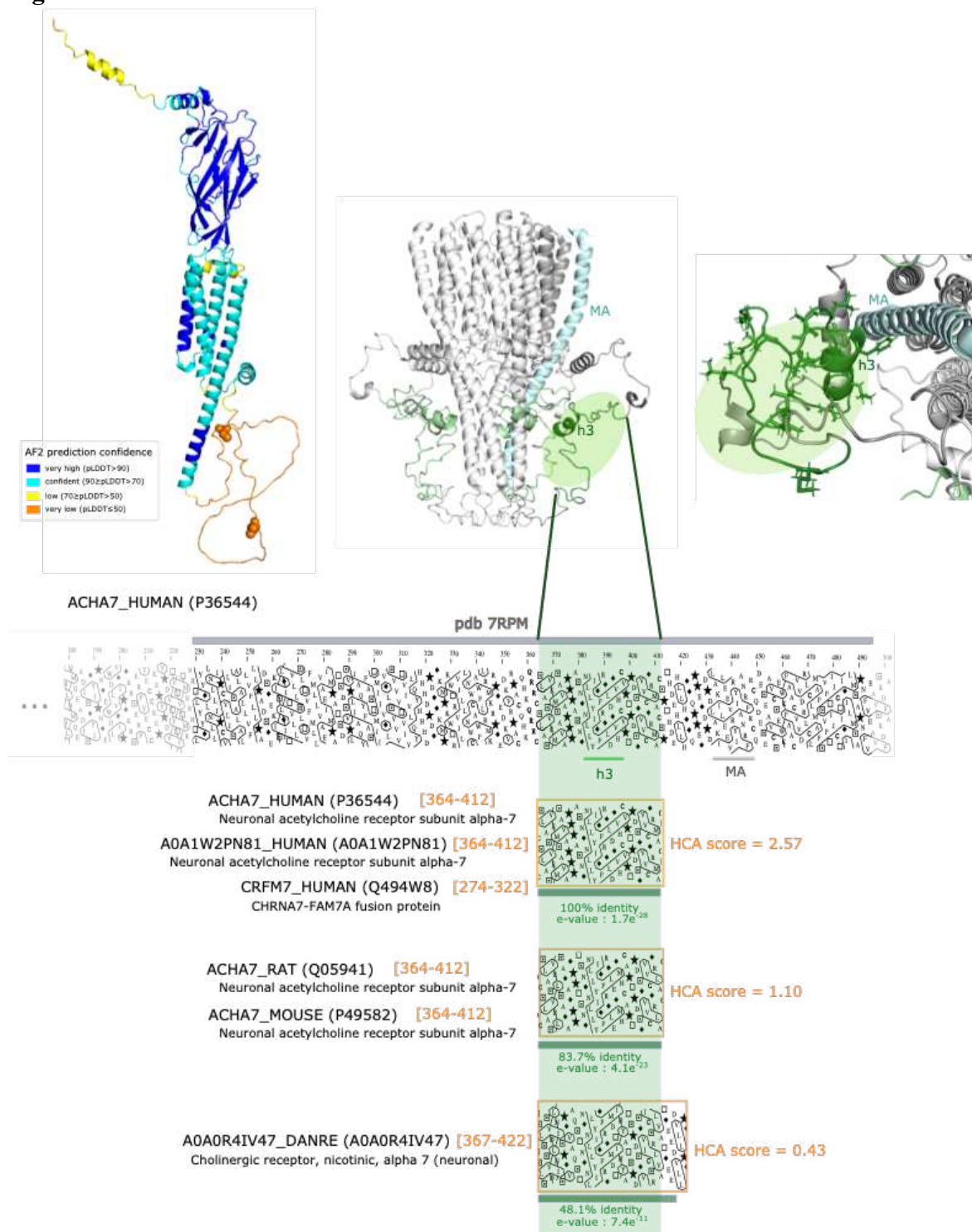
Figure S10



**Example of a long, disordered sequence, corresponding to a foldable segment with globular-like features (C-terminal transactivation domain of the CBP/p300-interacting transactivator 2 (CITED2)).**

The foldable segment shares the characteristics of a globular-like domain, as assessed by an HCA score of 0.98. At left is shown its experimental 3D structure (ribbon representation), in complex with the cysteine-histidine-rich 1 domain of p300/CBP (grey surface) (pdb 1P4Q). Of note on the HCA plot is the long cluster (aa 217 to 243), unusual in stable globular domains, in which loops linking the regular secondary structures are less hydrophobic. At right is shown the AlphaFold2 3D structure model of CITED2, colored according to the AlphaFold2 confidence score and superimposed on the experimental 3D structure (of which only the p300/CBP surface is shown). The AF2 model captures some elements of the bound conformation (blue shaded on the HCA plot - labels 2 to 4), while providing an additional helix at the C-terminal end (5), whose position is however not allowed in the CITED2/CBP complex. An additional helix upstream of the domain (aa 112 to 125) is predicted by AlphaFold2 (corresponding to a short foldable segment (HCA score -0.9, aa 114 to 125)), also in position not allowed in the CITED2/CBP complex.

Figure S11



**Example of a foldable segment, with AF2 very low confidence values over its entire length, and which was afterwards observed in a different conformation in an experimental 3D structure.**

This segment, belonging to the neuronal acetylcholine receptor subunit alpha-7, is highlighted on the AF2 3D structure model (on the left), as well as on the experimental 3D structure (in the middle and on the right, in green) within the oligomeric architecture of the protein (pdb 7RPM). The focus on the right shows the hydrophobic amino acids included in the hydrophobic clusters. At the bottom are shown the HCA plots of the foldable segments of the corresponding sequences, and related ones.

## SI references

1. Lamiable A, Bitard-Feildel T, Rebehmed J, et al. A topology-based investigation of protein interaction sites using Hydrophobic Cluster Analysis. *Biochimie*. 2019;167:68-80.
2. Hennetin J, Le Tuan K, Canard L, Colloc'h N, Mornon J-P, 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.
3. Callebaut I, Labesse G, Durand P, et al. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol Life Sci*. 1997;53:621-645.
4. Gaboriaud C, Bissery V, Benchetrit T, Mornon J-P. Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. *FEBS Lett*. 1987;224(1):149-155.
5. Eudes R, Le Tuan K, Delettré J, Mornon J-P, Callebaut I. A generalized analysis of hydrophobic and loop clusters within globular protein sequences. *BMC Struct Biol*. 2007;7:2.
6. Woodcock S, Mornon J-P, Henrissat B. Detection of secondary structure elements in proteins by hydrophobic cluster analysis. *Protein Eng*. 1992;5:629-635.
7. Faure G, Callebaut I. Comprehensive repertoire of foldable regions within whole genomes. *PLOS Comput Biol*. 2013; 9(10):e1003280.
8. Bitard-Feildel T, Callebaut I. Exploring the dark foldable proteome by considering hydrophobic amino acids topology. *Sci Rep*. 2017;7:41425.
9. Bitard-Feildel T, Lamiable A, Mornon J-P, Callebaut I. Order in disorder as observed by the "Hydrophobic Cluster Analysis" of protein sequences. *Proteomics*. 2018;18:e1800054.
10. Faure G, Callebaut I. Identification of hidden relationships from the coupling of hydrophobic cluster analysis and domain architecture information. *Bioinformatics*. 2013;29:1726-1733.
11. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2020;49(D1):D480-D489.
12. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods*. 2011;9:173-175.
13. Blum M, Chang H-Y, Chuguransky S, et al. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res*. 2020;49(D1):D344-D354.
14. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389-3402.
15. Lu S, Wang J, Chitsaz F, et al. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res*. 2020;48:D265-D268.
16. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*. 2016 33:1635-1638.
17. Sá-Moura B, Simões AM, Fraga J, et al. Biochemical and biophysical characterization of recombinant yeast proteasome maturation factor ump1. *Comput Struct Biotechnol J*. 2013;7 e201304006.
18. Uekusa Y, Okawa K, Yagi-Utsumi M, et al. Backbone <sup>1</sup>H, <sup>13</sup>C and <sup>15</sup>N assignments of yeast Ump1, an intrinsically disordered protein that functions as a proteasome assembly chaperone. *Biomol NMR Assign*. 2014;8:383-386.

19. Schnell HM, Walsh RMJ, Rawson S, et al. Structures of chaperone-associated assembly intermediates reveal coordinated mechanisms of proteasome biogenesis. *Nat Struct Mol Biol.* 2021;28:418-425.
20. Campbell EA, Masuda S, Sun J, et al. Crystal structure of the *Bacillus stearothermophilus* anti-sigma factor SpoIIAB with the sporulation sigma factor sigmaF. *Cell.* 2002;108:795-807.
21. Cartagena AJ, Banta AB, Sathyan N, et al. Structural basis for transcription activation by Crl through tethering of  $\sigma$ S and RNA polymerase. *Proc Natl Acad Sci U S A.* 2019;116:18923-18927.
22. Campen A, Williams RM, Brown CJ, Meng J, Uversky VN, Dunker AK. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept Lett.* 2008;15(9):956-963.







## **Chapitre 3. EXPLORATION DE L'INCONNU D'ALPHAFOLD2 : DU DESORDRE ET PLUS**

## Présentation de l'article

L'article présenté dans ce chapitre a été publié dans *Biomolecules*.

En combinant pyHCA et AlphaFold2 (AF2) dans l'étude présentée dans le chapitre précédent, nous avons pu mettre en avant une portion non négligeable de résidus dont les prédictions AF2 sont de très faible confiance, et qui sont situés dans des segments foldables typiques des domaines globulaires solubles (score HCA dans l'intervalle [-1,3.5]). Dans ce chapitre, nous avons souhaité explorer plus avant ces segments, et plus précisément, les segments foldables typiques de domaines globulaires solubles longs (taille > 30 acides aminés), et entièrement couverts par des résidus dont la prédiction AF2 est de très faible confiance ( $pLDDT \leq 50$ ), nous les avons notés VL pour *Very Low*. Deux possibilités peuvent expliquer la faible confiance avec laquelle sont prédits ces segments : (i) ils correspondent à de l'ordre conditionnel, c'est-à-dire, à des séquences réellement désordonnées et capables de se replier sous contrainte (e.g. interaction à un partenaire) ; (ii) ces segments correspondent à de « l'ordre caché » (structures stables, ordre non conditionnel), qu'AF2 ne peut détecter par absence d'information évolutive ou de structure similaire dans la banque de données d'apprentissage. Ces derniers segments représentent des cibles intéressantes pour une caractérisation expérimentale.

L'exploration des caractéristiques de ces segments associés à des scores pLDDT faibles est réalisée par comparaison à des segments sélectionnés sur les mêmes critères (segments foldables longs, avec des scores dans l'intervalle [-1,3.5]), mais qui sont entièrement couverts par des résidus pour lesquels les scores pLDDT sont très élevés (notés VH pour *Very High*).

Les segments VH et VL sont décrits par des caractéristiques déduites de leurs séquences, à savoir : le pourcentage de résidus prédits désordonnés par IUPred2 (Mészáros et al., 2018) et le nombre moyen par position d'homologues identifiés dans la banque de données BFD (Steinegger et al., 2019b; Steinegger et Söding, 2018). Des caractéristiques sont également extraites des structures 3D prédites par AF2 : le pourcentage d'acides aminés inclus dans des structures secondaires et le pourcentage d'acides aminés accessibles au solvant, ces deux informations étant extraites avec l'outil DSSP (Kabsch et Sander, 1983). Les segments VH et VL ont été classés suivant ces caractéristiques en utilisant un arbre binaire schématisé en Figure 3.1.

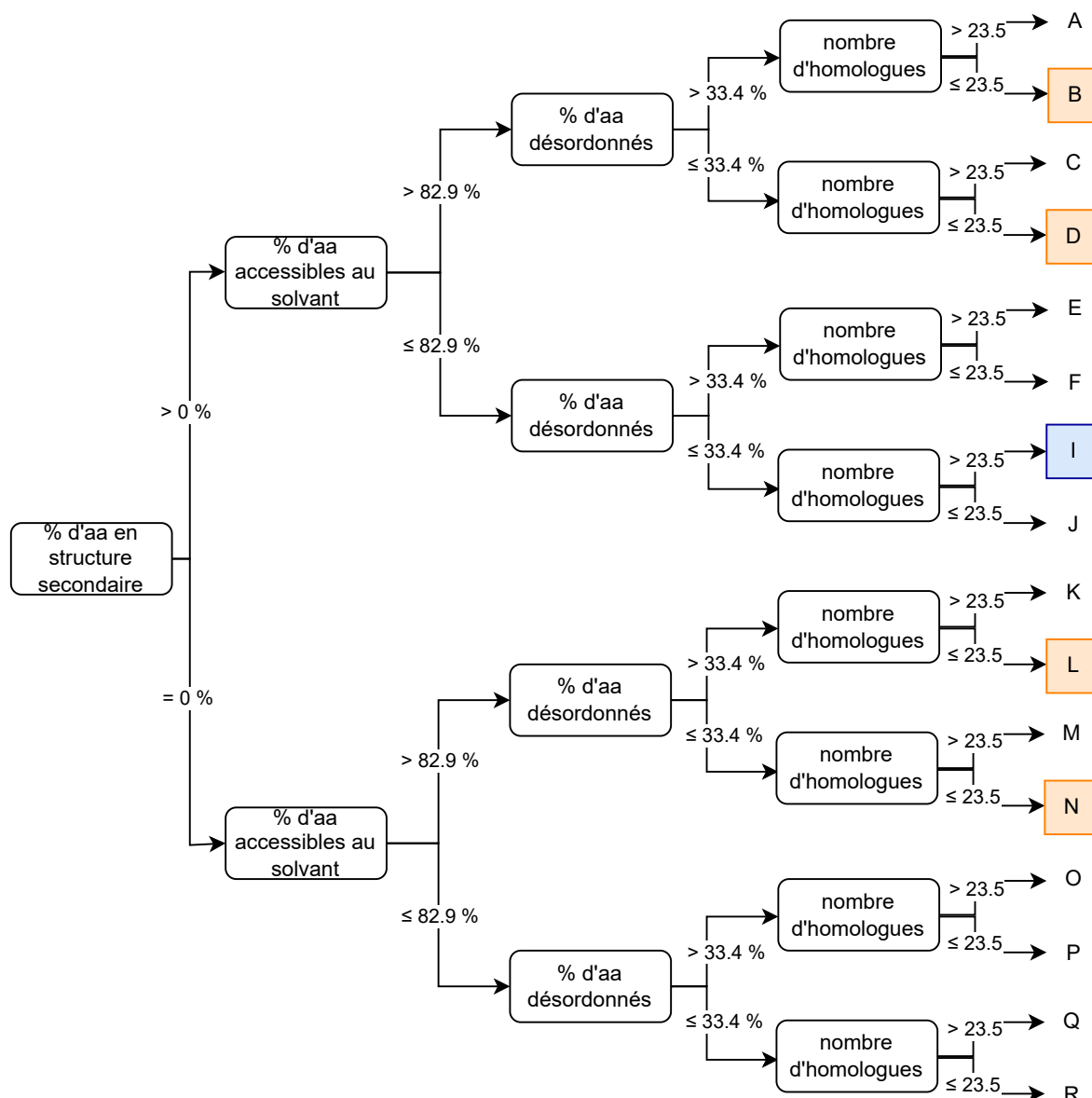


Figure 3.1. Arbre binaire utilisé comme support de classification des segments foldables de type globulaire soluble de plus de 30 acides aminés et dont les prédictions par AF2 sont VH ou VL (deux classification distinctes). Tous les seuils sauf celui du pourcentage en acides aminés des structures secondaires régulières ont été fixés pour regrouper les 95% des segments VH (5<sup>ème</sup> percentile pour l'accessibilité au solvant et le désordre et 95<sup>ème</sup> pour le nombre d'homologues). En bleu est encadré la catégorie majoritaire des segments VH (>90% des segments) et en orange les catégories majoritaires du VL (>10% pas catégorie) (aa = acides aminés). Les effectifs pour chaque catégorie sont représentés dans l'article en Figure 3 pour le VH et en Figure 5 pour le VL.

À l'exception de deux cas, tous les segments foldables globulaires solubles VH sont prédits par AF2 avec des structures secondaires régulières. La catégorie de segments foldables globulaires solubles longs VH la plus abondante (plus de 90% des segments) regroupe des modèles de structure 3D compacts (avec structures secondaires et faible accessibilité au solvant) et des séquences non ou peu désordonnées, avec un nombre important d'homologues dans BFD (encadré bleu en Figure 3.1). Les segments VL sont plus dispersés, quatre catégories

(encadrés orange en Figure 3.1) regroupent plus de 10% des segments. Dans l'article présenté ci-après, chacune de ces catégories est illustrée par des exemples (Figures 4, 7 et 8).

La majorité des segments de type globulaire soluble longs VL sont prédits sans structure secondaire régulière, avec une majorité de résidus accessibles au solvant, désordonnés ou non et avec peu d'homologues connus (catégories N et L Figure 3.1). Deux types de segments sont observés dans ces catégories, les premiers correspondant à de l'ordre conditionnel, hypothèse appuyée par leur annotation par la banque de données du désordre DisProt (structures secondaires transitoires) (Hatos et al., 2020). Les seconds, pour lesquels aucun désordre n'est prédit par IUPred2 et incluant des amas hydrophobes typiquement observés dans des domaines globulaires de la banque SCOPe, pourraient correspondre à des structures stables (ordre caché). L'échec d'AF2 peut s'expliquer par un manque d'information évolutive ou du fait que la structure corresponde à une nouvelle conformation, absente des banques de données d'AF2. Différents exemples illustrent ces cas de figure dans l'article présenté ci-après.

Pour valider la pertinence de pyHCA vis-à-vis de l'identification d'ordre dans les prédictions VL d'AF2, nous avons effectué une recherche de similitude entre ces longs segments foldables de type globulaire soluble VL et les séquences PDB dont les structures expérimentales ont été déterminées après le dépôt des prédictions dans AFDB (et donc non considérées dans l'apprentissage par AF2). Nous avons ainsi pu mettre en évidence une prédiction d'AF2 sans structure secondaire régulière et avec un taux d'acides aminés accessibles au solvant très élevé, pour lequel la structure déterminée expérimentalement par RMN est maintenant disponible (identifiant PDB : 7RPM, récepteur nicotinique acétylcholine alpha7 humain (Bondarenko et al., 2022)). Les auteurs de cet article montrent que si la structure 3D de ce récepteur présente une plasticité conformationnelle, elle adopte une conformation repliée dans son état de repos. Le segment foldable prédit avec un très faible score pLDDT contient trois amas hydrophobes, dont un correspondant à l'hélice h3 observée expérimentalement (Figure 3.2). Cet exemple appuie l'apport de pyHCA dans le contexte des prédictions AF2. Plusieurs segments foldables peuvent ainsi être mis en évidence, constituant des cibles de choix pour de futures études expérimentales en vue de caractériser leurs structures et fonctions.

Nous observons également une quantité importante de segments pour lesquels AF2 prédit des structures 3D compactes, avec un ensemble de structures secondaires régulières en interaction (% de résidus en structures secondaires > 0 et % de résidus accessibles < 82.9% ; catégories B et D Figure 3.1). En regard des tracés HCA, nous observons une bonne correspondance entre les structures secondaires prédites par AF2 et les amas hydrophobes. Aussi, on peut émettre l'hypothèse que la structure prédite par AF2 soit correcte, mais que le score reste trop faible à cause du manque d'information évolutive. Les structures secondaires peuvent également être plus ou moins bien prédites, mais mal assemblées.

Ces segments, dans lesquels pyHCA et AF2 détectent de l'ordre, sans que ce dernier ne parvienne à prédire leur structure 3D avec bonne confiance, seraient des cibles privilégiées à l'expérimentation.

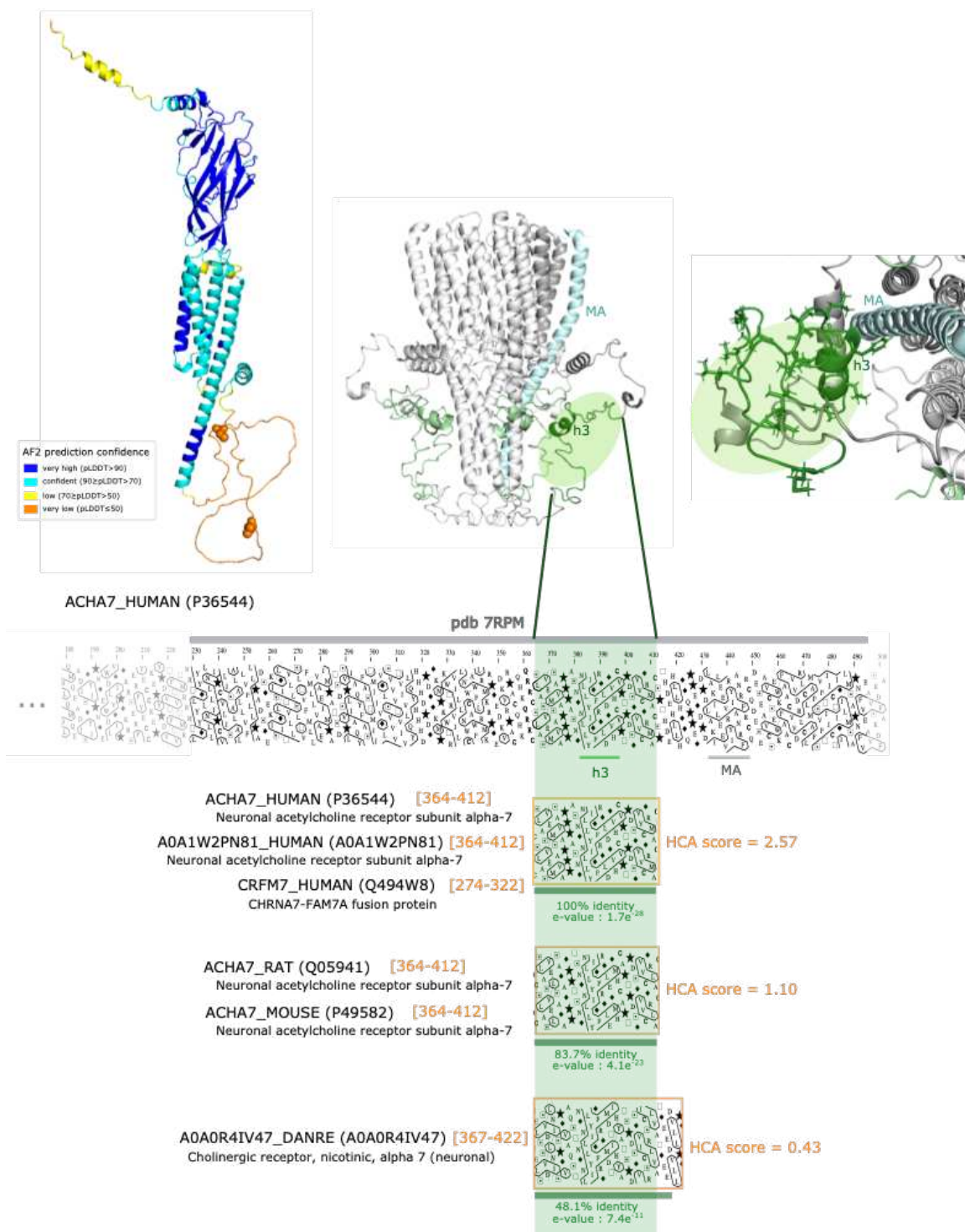



Figure 3.2. Exemple d'un segment foldable prédit par AF2 avec un score pLDDT très faible, et qui a ensuite été observé dans une conformation différente dans une structure 3D expérimentale. Ce segment, appartient à la sous-unité alpha-7 du récepteur neuronal de l'acétylcholine, est mis en évidence sur le modèle de structure 3D AF2 (à gauche), ainsi que sur la structure 3D expérimentale (au milieu et à droite, en vert) au sein de l'architecture oligomérique de la protéine (pdb 7RPM). Dans le focus sur la droite les acides aminés hydrophobes inclus dans les clusters hydrophobes sont mis en évidence sous forme de bâtonnets. En bas, sont montrés les diagrammes HCA des segments foldables VL présentant des similitudes de séquence avec 7rpm.

## Article

# Digging into the 3D Structure Predictions of AlphaFold2 with Low Confidence: Disorder and Beyond

Apolline Bruley, Jean-Paul Mornon, Elodie Duprat <sup>\*,†</sup>  and Isabelle Callebaut <sup>\*,†</sup>

Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, 75005 Paris, France

\* Correspondence: elodie.duprat@sorbonne-universite.fr (E.D.); isabelle.callebaut@sorbonne-universite.fr (I.C.)

† These authors contributed equally to this work.

**Abstract:** AlphaFold2 (AF2) has created a breakthrough in biology by providing three-dimensional structure models for whole-proteome sequences, with unprecedented levels of accuracy. In addition, the AF2 pLDDT score, related to the model confidence, has been shown to provide a good measure of residue-wise disorder. Here, we combined AF2 predictions with pyHCA, a tool we previously developed to identify foldable segments and estimate their order/disorder ratio, from a single protein sequence. We focused our analysis on the AF2 predictions available for 21 reference proteomes (AFDB v1), in particular on their long foldable segments (>30 amino acids) that exhibit characteristics of soluble domains, as estimated by pyHCA. Among these segments, we provided a global analysis of those with very low pLDDT values along their entire length and compared their characteristics to those of segments with very high pLDDT values. We highlighted cases containing conditional order, as well as cases that could form well-folded structures but escape the AF2 prediction due to a shallow multiple sequence alignment and/or undocumented structure or fold. AF2 and pyHCA can therefore be advantageously combined to unravel cryptic structural features in whole proteomes and to refine predictions for different flavors of disorder.

**Keywords:** long foldable segments; pyHCA; soluble domains; protein sequence; conditional order; hidden order; dark proteomes; intrinsically disordered domains



**Citation:** Bruley, A.; Mornon, J.-P.; Duprat, E.; Callebaut, I. Digging into the 3D Structure Predictions of AlphaFold2 with Low Confidence: Disorder and Beyond. *Biomolecules* **2022**, *12*, 1467. <https://doi.org/10.3390/biom12101467>

Academic Editors: Stefania Brocca, Keith Dunker, Sonia Longhi and Prakash Kulkarni

Received: 14 September 2022

Accepted: 5 October 2022

Published: 13 October 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

AlphaFold2 [1] and RoseTTAfold [2] have recently achieved an impressive breakthrough in the field of structural biology, providing accurate models of three-dimensional (3D) structures of proteins based on only knowledge of their amino acid sequences alone. Based on deep-learning techniques, they take advantage of the vast existing knowledge of protein sequences and 3D structures, recently expanded through environmental genomics and structural genomics approaches. In particular, they extensively used evolutionary information to detect co-variation of residues (or correlated mutations), the underlying idea being that residues that have co-evolved are close in 3D space. The first version of the AlphaFold2 database (AFDB v1) [3] included predictions for a very large part of proteomes from 21 widely studied organisms. It has been extended to provide open access to over 200 million predictions, covering nearly every organism with protein sequence data. This provides the scientific community with a wealth of knowledge, which could accelerate the understanding of protein structure-function relationships and have a profound impact on many areas of biology, including human health and the environment.

Several studies have already been conducted to estimate the extent to which AlphaFold2 (AF2) improves the coverage in structural biology, as well as to analyze its current advantages and limitations (e.g., [4–10]). One striking feature of AF2 is that it provides a per-residue metric, reflecting confidence in the structural assignment (predicted local distance difference test (pLDDT)) [1]. High values of pLDDT are observed for folded



domains, contrasting with low values typically associated with linkers and unstructured or disordered regions [11]. The relevance of pLDDT as a predictor of disorder has been supported on the CAID benchmark dataset [12] and compared to other state-of-the-art disorder predictors, such as SPOT-Disorder2 or IUPred2 [1,4,11].

At least two questions need to be considered when focusing on very low confidence regions (pLDDT < 50) in AF2 predictions, which are assumed to be globally disordered. The first question is whether it is possible to reveal conditional order within these intrinsically disordered regions (IDRs), from amino acid sequence information alone. Such IDRs may be involved in molecular recognition, to which hydrophobic interactions make major contributions [13,14]. In many cases, these regions undergo a disorder-to-order transition (induced folding) to a more structured state upon binding with a partner [15]. High-resolution multi-dimensional NMR studies have demonstrated that such IDRs, ranging in length from 10 up to 70 amino acids and referred to over time by different names ([16,17], molecular recognition elements [18], primary contact sites [19], preformed structural elements [20], pre-structured motifs [21]), can be pre-populated by transient local structural elements, presaging the target-bound conformation [21]. The plasticity of these IDRs can allow for a range of secondary structures in the bound state, as shown by the example of the p53 tumor suppressor protein [22]. Some IDRs are also able to retain a significant degree of structural heterogeneity in the bound states [23], leading to the definition of fuzzy complexes [24,25]. Some IDRs involved in molecular recognition consist of or incorporate short linear motifs (SLiMs), i.e., short conserved sequences, which enable low affinity, transient, and conditional interactions and are often located within disordered regions [26]. Specifying the structural unit in which these short interacting motifs are embedded should inform on the global features of the interaction (such as affinity, specificity, fuzziness). Regarding conditional order, another question to consider is whether it is possible to identify, within these very low confidence AF2 predictions, longer binding IDRs that meet the definition of intrinsically disordered domains (IDD) [27–29], which must be stabilized by a partner within protein complexes to adopt a stable fold?

The second issue related to very low pLDDT regions is to evaluate whether some might not be disordered and might still adopt a well-folded 3D structure, but AF2 cannot predict it (what we call “hidden order”). This hypothesis is conceivable as the co-evolutionary information, necessary to predict inter-residue contacts, is lacking for some protein sequences. These proteins, not predicted as disordered, escape any annotation coming from sequence or structure databases and constitute the dark proteome [30,31]. They still represent 10% of the human proteome after annotation with the AlphaFold2 predictions [7].

We recently analyzed the pLDDT values observed for the AF2 3D structure predictions on the 21 reference proteomes (AFDB v1) in light of another metric, called the HCA score (Bruley et al. [32]). The HCA score is based on Hydrophobic Cluster Analysis (HCA), a two-dimensional approach allowing the analysis of the content of an amino acid sequence in regular secondary structures (see [33] for a recent review of the methodology). Indeed, the hydrophobic clusters defined by this approach mainly correspond to the positions of regular secondary structures constituting the building blocks of folded domains [34–36]. The analysis of the composition of a sequence in hydrophobic clusters thus provides information on its architecture in domains and the disorder/order content of the delineated domains. A tool has been developed to automatically partition protein sequences into foldable segments based on a measure of hydrophobic cluster density [37]. The calculation of an HCA score provides information about the composition of the sequence in clusters and hydrophobic amino acids within the clusters, which thus reflects the overall order/disorder ratio of the foldable segments [32]. Using this HCA score, we disentangled different types of disorder and appreciate disorder-to-order continuum. While residues with low-pLDDT values were enriched in non-foldable segments, a significant portion of foldable segments with HCA scores typical of well-folded domains also had low mean pLDDT values in AF2 3D structure predictions. This suggests that these regions carry specific functional information

(corresponding to the two cases mentioned above) that remains unraveled by AF2 (Bruley et al. [32]).

Here, we further explored the source of this apparent inconsistency between foldability and low confidence AF2 prediction, which is widely assimilated to disorder in the literature. To this end, we analyzed, from the same 21 reference proteomes, the long soluble-like foldable segments as defined by the pyHCA tool, whose residues all have a very low AF2 pLDDT value (hereafter referred to as full-VL segments). Particular focus was on segments of length >30 amino acids, which corresponds to the minimum length considered for globular domains [38]. Moreover, this minimal length excludes a large number of short motifs (SLiMs) undergoing induced folding and which are otherwise associated with higher HCA score values (Bruley et al. [32]). To analyze these long, soluble-like full-VL segments, we considered four features related to their amino acid sequences and 3D structures, as predicted by AF2. We described these 3D structure models by the proportion of residues involved in a regular secondary structure (RSS) and by the proportion of residues accessible to the solvent. In addition, we described the protein sequences on which these predictions were based, by the proportion of residues predicted as disordered by IUPred2 and by the average number of homologs per residue as found in the large environmental BFD database. The latter feature allowed us to consider co-evolution information, essential for the reliable prediction of amino acid contacts by AF2. We compared these features to their distribution for long soluble-like foldable segments whose residues all have a very high pLDDT values (hereafter referred to as full-VH segments), for the 21 proteomes included in AFDB v1.

## 2. Material and Methods

### 2.1. Proteomes from AlphaFold Protein Structure Database v1.

Amino acid (aa) sequences and predicted 3D structures were downloaded from the AlphaFold Protein Structure database (AFDB) v1 ([3], <https://alphafold.ebi.ac.uk>, accessed on 21 July 2021) for the 21 reference model-organism proteomes. The per-residue model confidence values (pLDDT) were extracted from the 3D coordinate files (B-factor column in PDB format).

### 2.2. Delineation of Soluble-like Foldable Segments within Protein Sequences

The *segment* function of the pyHCA tool (provided at <https://github.com/DarkVador-HCA/pyHCA>) was used to automatically delineate foldable segments (FS), i.e., segments with a high density of hydrophobic clusters (HC), as defined by the Hydrophobic Cluster Analysis (HCA) [33]. HC consist of strong hydrophobic amino acids (V,I,L,M,F,Y,W) and are separated from each other by at least four other amino acids or a proline. For FS delineation, cysteine (C) is integrated into the hydrophobic alphabet and HC consist of only one or two consecutive hydrophobic amino acid(s) are not considered, as they are mainly associated with coils [36].

The HCA score, which measures the density of hydrophobic clusters and strong hydrophobic amino acids of foldable segments (Bruley et al. [32]), was calculated using the *segment* function (pyHCA tool). Soluble-like segments were defined according to an HCA score value between  $-1$  and  $3.5$ .

### 2.3. Description of Sequence and Structural Features

Our final dataset consisted of proteins from AFDB v1, encompassing at least one long (> 30 a.a.), globular soluble-like ( $-1 \leq \text{HCA score} \leq 3.5$ ) foldable segments, entirely made of residues with very low (VL) or very high (VH) 3D prediction confidence (pLDDT  $\leq 50$  and pLDDT  $> 90$ , respectively). We considered four different features to characterize the amino acid sequence and AF2 3D models of these segments, as detailed below. For each feature, we defined a threshold value based on the distribution of these full-VH segments and delimiting an interval encompassing at least 95% of them. These threshold values were further used for the dataset description by binary trees.

### 2.3.1. Per-Residue Disorder Prediction

Disorder was predicted using the IUPred2A [39] *long* disorder predictor on the whole protein sequences. IUPred2A calculates a per-residue score between 0 and 1 that reflects the estimated stabilizing effect of other residues on each residue of one amino acid sequence. The coverage of the FS by disorder was then calculated (in percentage of the segment length), considered as disordered amino acids having a score above 0.5. The coverage threshold was set to 33.4%. The number of FS with a value below this threshold were as follows: 14,077 segments over 30,644 and 10,827 segments over 11,395 in case of full-VL and full-VH, respectively.

### 2.3.2. Known Homologs

The multiple sequence alignments used to build the AlphaFold2 models were not provided in AFDB repositories. Therefore, a search for known homologs in the reduced Big Fantastic Database (BFD) was performed using *jackhmmer* (from HMMER 3.3.2 [40], <http://hmmer.org/>). The parameters (e-value threshold of 0.0001, 1 iteration) were those used by AF2 in the similarity search step. The Big Fantastic Database (BFD) [1] (<https://bfd.mmseqs.com>, accessed on 24 May 2022) is a database containing 2.5 billion clustered protein sequences. It is the most comprehensive database used by AF2 in order to build multiple sequence alignments, gathering sequences from genomic and metagenomic databases (UniprotKB [41] and metaclust [42] and datasets assembled with Plass [43]). The reduced version of BFD contains only representative sequences of each cluster (65,984,053 sequences). This one was downloaded following the recommendations given on the AF2 github (<https://github.com/deepmind/alphafold>, accessed on 24 May 2022). In this work, the sequence similarity search was performed on the whole protein sequences. The number of aligned sequences per FS position was then calculated and averaged over the length of the FS. The mean number threshold was set to 23.5 BFD homologs per segment residue. The number of FS with a value above this threshold were as follows: 3,347 segments over 30,644 and 10,829 segments over 11,395 in case of full-VL and full-VH, respectively.

### 2.3.3. Secondary Structure Assignment

Secondary structures were assigned from the coordinates of the AF2 3D structure models (PDB files, full-length proteins) using the DSSP program [44] available in the biopython module v1.78 for python v3.6.3. All amino acids found in alpha helices (encoded as “H” in DSSP), 3–10 helices (“G”), Pi helices (“I”), strands (“E”), and isolated beta-bridge residues (“B”) were considered to participate in regular secondary structures (RSS). The percentage of the FS residues participating in a RSS was then calculated. The number of FS with at least 1 RSS were as follows: 10,993 segments out of 30,644 and 11,393 segments out of 11,395 in case of full-VL and full-VH, respectively.

### 2.3.4. Solvent Accessibility

Using the same module, the residues relative accessible surface area was calculated. This value was obtained by normalizing the residue accessible surface area (ASA) by the maximum ASA for the residue, computed on Gly-X-Gly tripeptides (where X is the residue of interest). By default, DSSP referred to the Sander and Rost scale for maximum ASA values per residue [45]. We considered a residue to be solvent accessible if the relative ASA was above 0.36 (based on Rost and Sander [45]). The percentage of accessible residues was calculated on each FS. The feature threshold was set to 82.9%. The number of FS with a value below this threshold were as follows: 1,908 segments over 30,644 and 10,823 segments over 11,395 in case of full-VL and full-VH, respectively.

### 2.3.5. 3D Structure Comparison

The Dali server ([46], <http://ekhidna2.biocenter.helsinki.fi/dali>) was used to compare the AF2 3D structure models of the foldable segments with PDB experimental 3D structures.

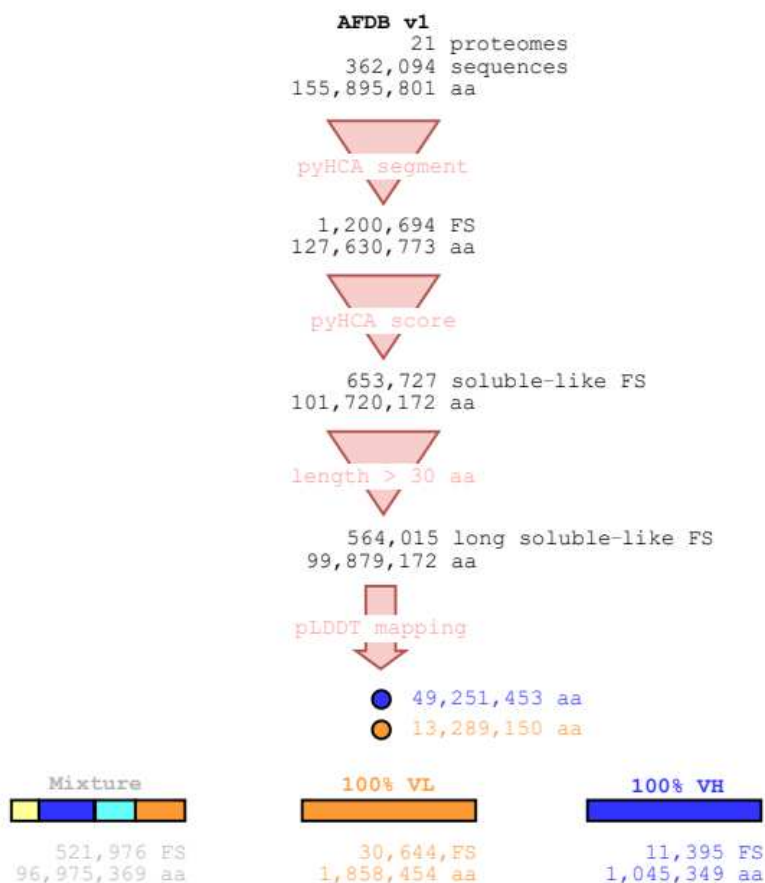
### 2.3.6. Figure Creation

3D structures were visualized with the UCSF Chimera software [47]. HCA plots were drawn using the DrawHCA program (<http://osbornite.impmc.upmc.fr/hca/hca-seq.html>). Hydrophobic clusters (HC) affinities for RSS were extracted from HCDB [36]. Binary tree diagrams were created using the R package *ggparty* (<https://github.com/martin-borkovec/ggparty>).

## 3. Results

### 3.1. General Features of Full-VL and Full-VH Segments from AFDB v1

Figure 1 illustrates the technical flow used in this study to extract 30,644 full-VL and 11,395 full-VH long soluble-like foldable segments from AFDB v1 using the pyHCA tool.

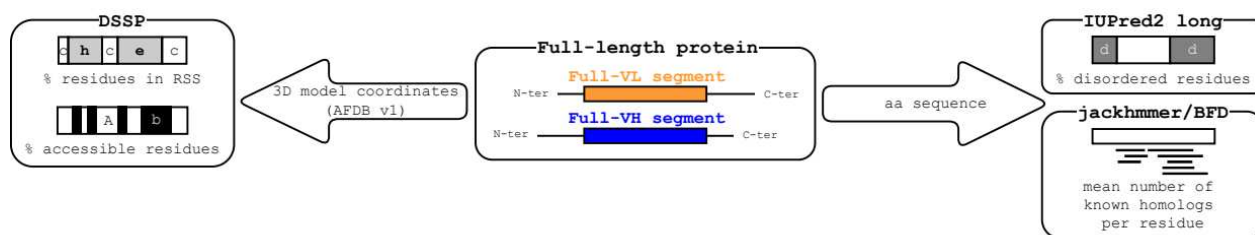


**Figure 1.** The technical flow for definition of the long soluble-like full-VH and full-VL foldable segments from AFDB v1 by using the pyHCA tool. The number of foldable segments (FS) and the number of residues (aa) are indicated at each step of the flow. The dataset further analyzed in this study consists of the full-VL and full-VH segments. For quantitative details about each of the 21 proteomes, see Supplementary Table S1.

Details for each of the 21 proteomes are given in Supplementary Table S1. Most of the residues in AFDB v1 (64.1%) are included in long soluble-like foldable segments (from 55.5% up to 73.9% in the proteomes of *Leishmania infantum* and *E. coli*, respectively). These segments are mainly composed of residues with a very high pLDDT value (49.3% VH, 13.3% VL). This trend is

also observed for each proteome, except for *Plasmodium falciparum* (23.8% VH, 40.4% VL). However, the set of the full-VL segments is larger than the set of full-VH segments, both in the number of segments and in the number of residues (Figure 1). This trend is observed for each of the 17 eukaryotic proteomes, where at least 9.1% VL residues included in a long soluble-like foldable segment are part of a full-VL segment (up to 22.4% and 24.0% for *Leishmania infantum* and *Plasmodium falciparum*, respectively). On the contrary, less than 6.3% of the VL residues included in a long soluble-like foldable segment are part of a full-VL segment for prokaryotic proteomes, where only a few cases of full-VL segments were found (1, 8, 14, and 29 segments for the archaeon *Methanocaldococcus jannaschii* and the bacteria *Escherichia coli*, *Staphylococcus aureus*, and *Mycobacterium tuberculosis* respectively). Furthermore, for eukaryotic proteomes, less than 2.6% of the VH residues included in a long soluble-like foldable segment are part of a full-VH segment (from 3.7% up to 8.1% for the prokaryotic proteomes). In AFDB v1, the mean length of full-VL segments (60.7 amino acids) is smaller than the mean length of full-VH segments (91.7 aa). This trend is observed for each of the 21 proteomes.

Figure 2 illustrates the technical flow used in this study for the description of the AF2 3D models and protein sequences for the full-VL and full-VH segment datasets. We described each segment by four quantitative features and explored their distribution for each dataset.

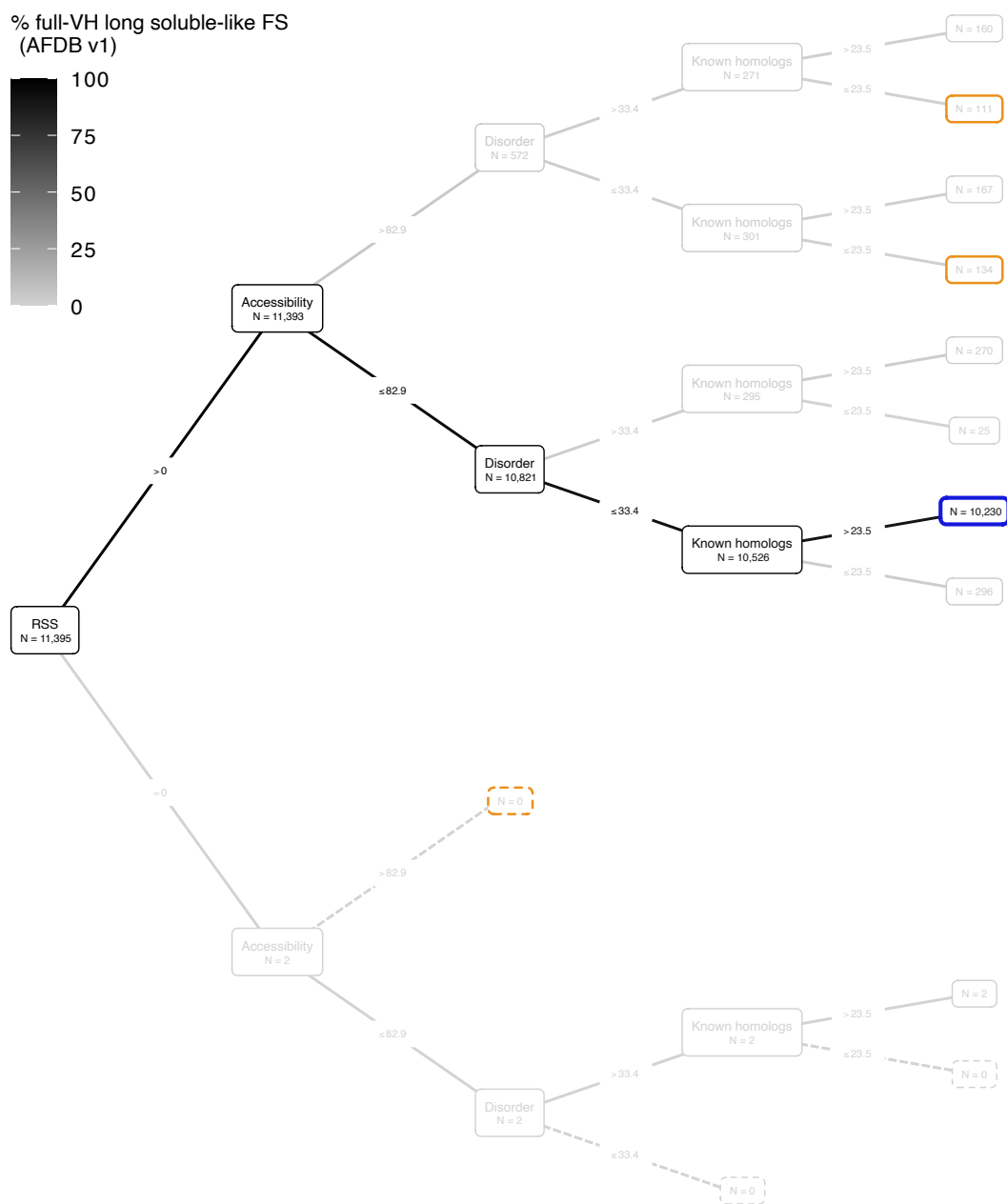


**Figure 2.** The technical flow for feature description of the segment dataset. Each AFDB v1 full-length protein comprising at least one full-VL or one full-VH long soluble-like foldable segment was analyzed by different tools (DSSP on the 3D coordinates, IUPred2 long, and jackhmmer on the amino acid sequence) allowing for calculation of four quantitative features describing each segment. Labels used for the different tools are: i) for DSSP secondary structure assignment: h, helix; e, strand (extended); c, coil; ii) for DSSP solvent accessibility: A, accessible, b, buried; iii) for IUPred2 long: d, disorder.

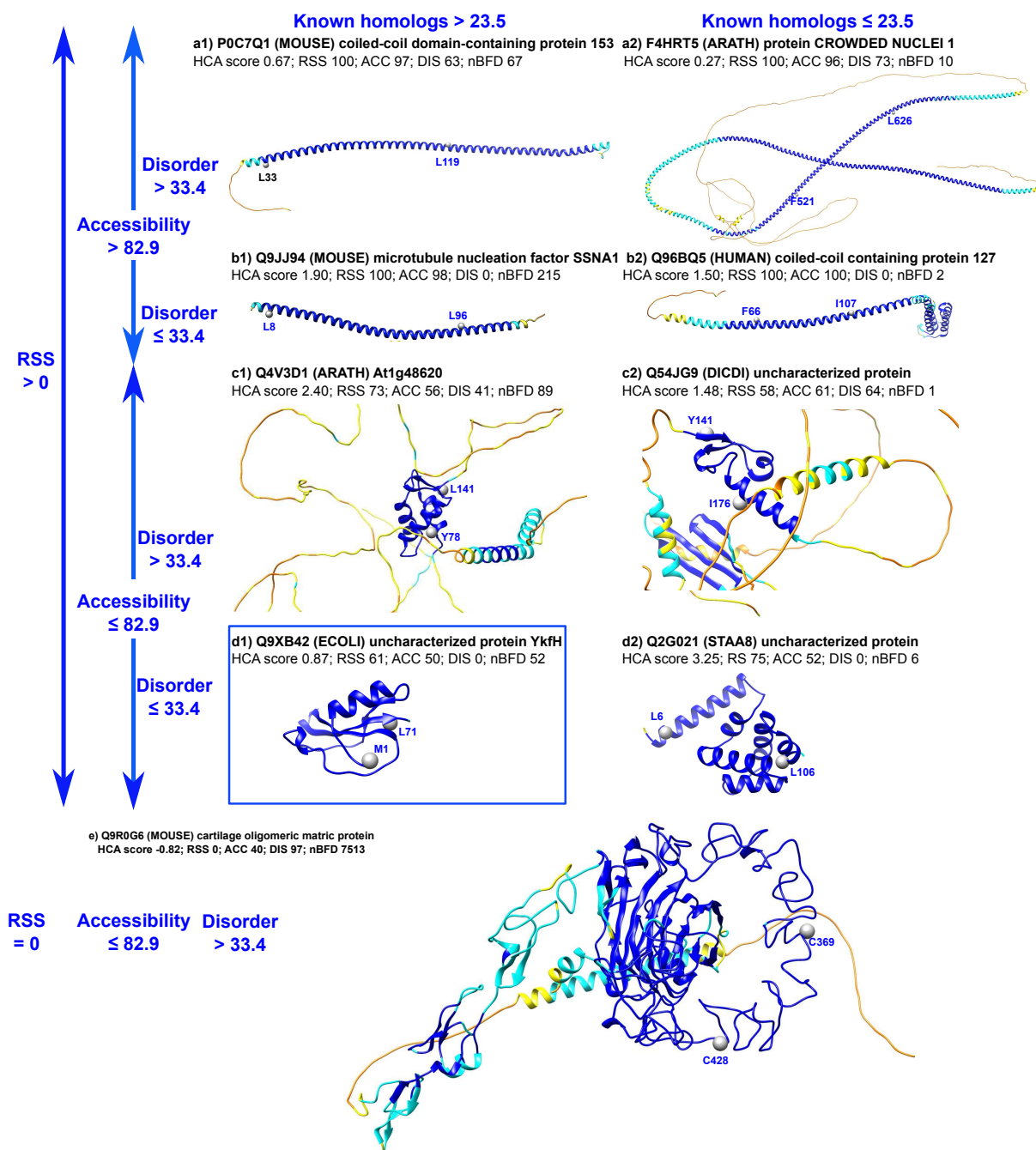
### 3.2. Full-VH Segments

Figure 3 depicts the classification of the 11,395 full-VH segments using a binary tree based on the features used to describe the 3D models and the amino acid sequences (see Figure 2 and Section 2 for details). Representative examples of the different categories are shown in Figure 4.

Quantitative thresholds were defined for each feature based on 95% full-VH, except for the proportion of segment residues participating in a RSS, as assigned by DSSP from the AF2 3D models (see Section 2 for details). For this 3D feature, we considered two classes of segments based on the presence/absence of RSS. All the long soluble-like foldable segments whose residues all have a very high pLDDT value (full-VH segments) are associated with the presence of RSS, except for two cases, corresponding to thrombospondin (TSP) repeats (Figure 4e). As observed in the experimental 3D structures that can serve as templates for homology modeling (pdb entries 1yo8 and 3fby), TSP repeats are folded domains with calcium ions bound into the core through acidic (aspartate) residues. The foldable segments delineated here contain conserved cysteine residues that form interdomain disulfide-bridges, providing tight interactions in the wire architecture typical of the TSP-2 signature domain [48].



**Figure 3.** Binary tree diagram of the full-VH segments according to the feature thresholds. The four levels of the tree (from the root on the left to the last internal nodes on the right) correspond to the four features describing the segments (see Figure 2 for the technical flow), as follows: percentage of segment residues participating in a regular secondary structure (RSS), percentage of segment residues accessible to the solvent (Accessibility), percentage of segment residues predicted to be disordered (Disorder), the mean number of BFD homologs per segment residue (Known homologs). The binary conditions based on each feature threshold are indicated on the edges of the tree (for details, see Section 2). The number of foldable segments with a given feature below or above each threshold is indicated in the internal and terminal nodes. The total number of full-VH segments is indicated within the root node. The terminal nodes corresponding to the most abundant subsets of full-VH segments (this figure) and full-VL segments (section 3.3.) are highlighted in blue and orange, respectively. For quantitative details about each of the 21 proteomes, see Supplementary Figure S1.



**Figure 4.** Examples of full-VH soluble-like foldable segments, distinguished according to the four features. The examples were extracted from the binary tree diagram shown in Figure 3. The AF2 3D structure models are colored according to pLDDT values, with the positions of the first and last amino acids of the full-VH soluble-like foldable segments indicated. The corresponding HCA score values are also reported, as well as those of the four features. The example extracted from the most populated leaf in Figure 3 is boxed in blue. HCA plots of the corresponding sequences are illustrated in Supplementary Figure S2. Subfigures a show examples with RSS > 0, Accessibility > 82.9, Disorder > 33.4 and BFD homologs per position > 23.5 (a1) and ≤ 23.5 (a2). Subfigures b show examples with RSS > 0, Accessibility > 82.9, Disorder ≤ 33.4 and BFD homologs per position > 23.5 (b1) and ≤ 23.5 (b2). Subfigures c show examples with RSS > 0, Accessibility ≤ 82.9, Disorder > 33.4 and BFD homologs per position > 23.5 (c1) and ≤ 23.5 (c2). Subfigures d show examples with RSS > 0, Accessibility ≤ 82.9, Disorder ≤ 33.4 and BFD homologs per position > 23.5 (d1) and ≤ 23.5 (d2). Subfigure 2e shows one of the two similar cases with RSS = 0, Accessibility ≤ 82.9, Disorder > 33.4.

The most abundant category of full-VH long soluble-like foldable segments (10,230 full-VH segments over 11,395, boxed in blue in Figures 3 and 4(d1)) corresponds to folded domains with low predicted disorder and a high number of BFD homologs. Domains were considered as folded as they contain RSS assembled together and have relative low solvent accessibility due to the involvement of a large number of amino acids in a hydrophobic core. Supplementary Figure S2 provides details of the HCA plots of the foldable segments whose 3D structures are shown in Figure 4. The folded domains contain  $\sim 1/3$  strong hydrophobic amino acids distributed in clusters, which correspond to the positions of RSS. A significant number of cases also exist with a smaller number of BFD homologs (296 segments, Figures 3 and 4(d2)). Here, the consideration of experimental 3D structures as templates can explain the accurate AF2 prediction (pdb:1sed for the example shown in Figure 4(d2)). Other interesting cases are those of folded domains corresponding to sequences predicted to be disordered for a large part, but which are clearly not (Figures 4(c1) and 4(c2) corresponding to histone fold, 29% identity with pdb 2lso-A, and to a case with no obvious similarity with known 3D structures, respectively). Finally, the cases of accurate AF2 predictions associated with models globally accessible to the solvent concern long helices, typical of coiled-coil assembly, whose sequences are predicted as disordered or not (Figure 4(a1,a2,b1,b2)). When no experimental 3D structure is available, the AF2 prediction is supported by a sufficiently informative periodic pattern and self-organizing structure, regardless of the number of BFD homologs.

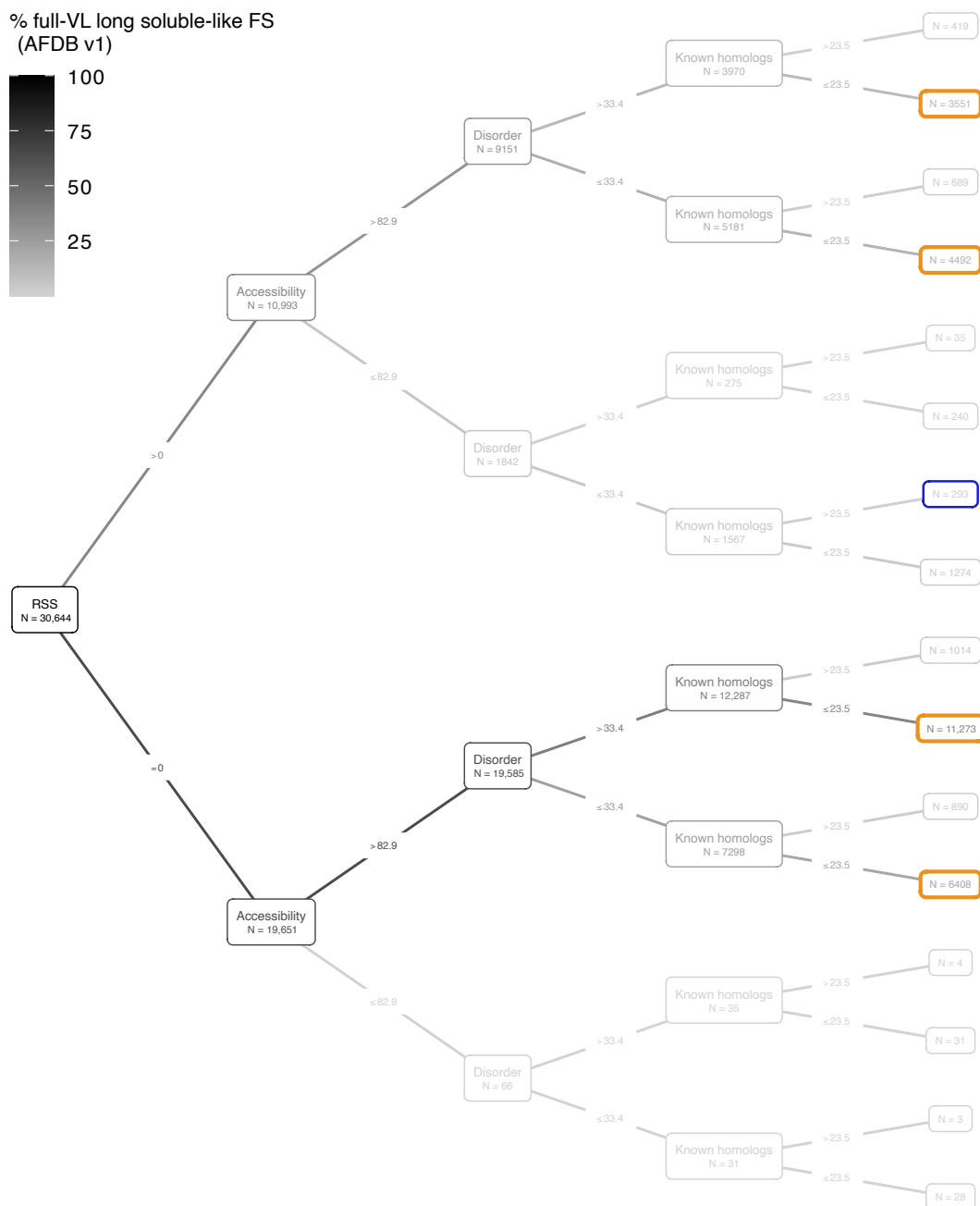
### 3.3. Full-VL Segments

Figure 5 shows the binary tree diagram of full-VL, long soluble-like foldable segments, according to the same threshold values as in Figure 3 for full-VH segments. The full-VL segments are much more dispersed across the different categories than the full-VH segments (see boxes in Figures 3 and 5). Four categories are populated by at least 10% of the full-VL segments. In contrast, there was only one in category in this case for full-VH segment, including 90% of them. Another notable point is that the mean values of the four features (RSS, Accessibility, Disorder, Known homologs) differ significantly between full-VH and full-VL segments, even when considering a same binary class (Figure 6). In particular, (i) full-VL segments with at least one RSS contain on average fewer residues participating in a RSS than similar full-VH segments (Figure 6a); (ii) full-VL segments with accessibility less than 82.9% are more accessible to solvent than similar full-VH segments (Figure 6b); (iii) full-VL segments with disorder less than 33.4% are predicted to be more disordered than similar full-VH segments (Figure 6c); finally, the full-VL segments with at least 23.5 known homologs per site in BFD have fewer homologs than similar full-VH segments (Figure 6d).

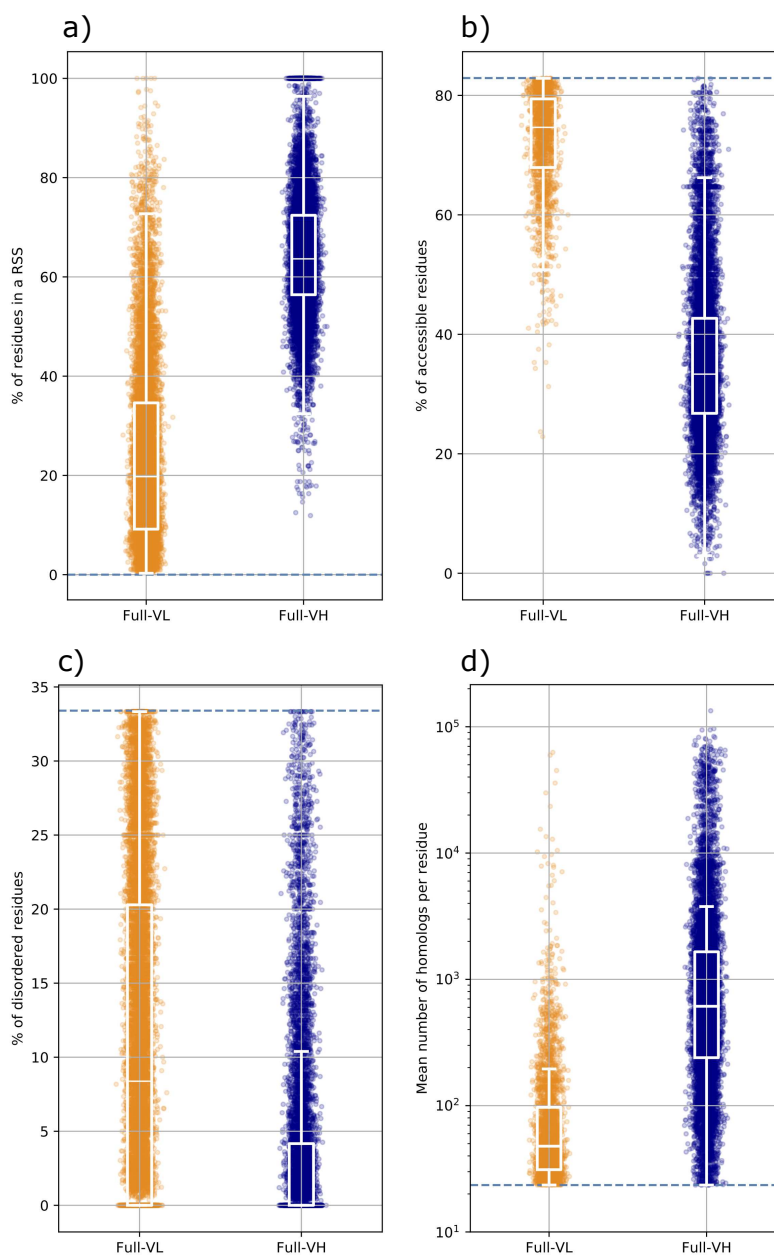
#### 3.3.1. Full-VL Segments with AF2 Well-folded Models

The category that is most populated for full-VH segments, i.e., 3D models with low solvent accessibility and tight contacts between the RSS, accounts for a substantial number of full-VL cases, although not predominant (293 segments: Figure 5, blue box). These AF2 predictions correspond to well-folded 3D structures, as illustrated with the yeast uncharacterized protein YBR032W (UniProt P38223, Figure 7b, blue box). This was predicted as an alpha + beta fold, but no significant structural similarity could be detected in the PDB database by the Dali server.



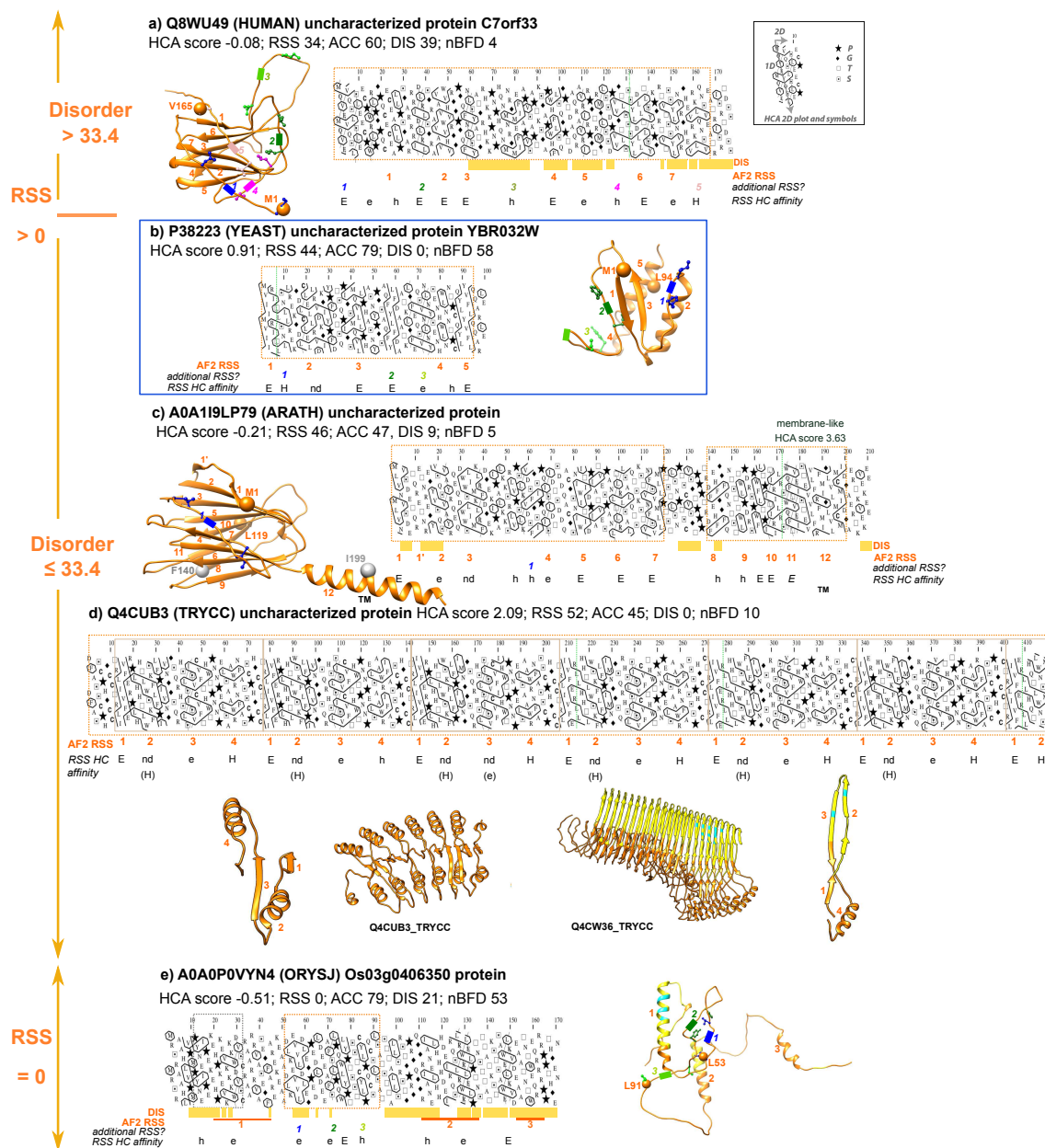


**Figure 5.** Binary tree diagram of the full-VL segments according to the feature thresholds. The four levels of the tree (from the root on the left to the last internal nodes on the right) correspond to the four features describing the segments (see Figure 2 for technical flow), as follows: percentage of segment residues participating in a regular secondary structure (RSS), percentage of segment residues accessible to the solvent (Accessibility), percentage of segment residues predicted to be disordered (Disorder), mean number of BFD homologs per segment residue (Known homologs). The binary conditions based on each feature threshold are indicated on the edges of the tree (for details, see Section 2). The number of foldable segments with a given feature below or above each threshold is indicated within the internal and terminal nodes. The total number of full-VL segments is indicated in the root node. The terminal nodes corresponding to the most abundant subsets of full-VL segments (this figure) and full-VH segments (Figure 3) are highlighted in orange and blue, respectively. For quantitative details about each of the 21 proteomes, see Supplementary Figure S3.



**Figure 6.** Distribution of features of 3D models and amino acid sequences for full-VH (blue) and full-VL (orange) long soluble-like foldable segments from AFDB v1 (21 proteomes). For the structural feature corresponding to the percentage of segment residues participating in a regular secondary structure (RSS) (a), only segments with at least 1 RSS as assigned by DSSP from the full-length protein 3D coordinates are shown (see Section 2 for quantitative details). For each feature in (b–d), the blue dashed line indicates the threshold value defined based on 95% of the full-VH segments (see Section 2 for details). For both full-VL and full-VH segments, only values falling in these intervals are shown.

Such AF2 predictions cannot be reported with high confidence for several reasons. They could correspond to the adopted structures, but represent novel folds, with amino acid contacts not yet described in the folds used for the AF2 machine learning step and insufficient depth of the multiple sequence alignment. Conversely, RSS could also be misassembled or insufficiently relative to what is happening in the actual structure.



**Figure 7.** Examples of full-VL soluble-like foldable segments corresponding to folded AF2 predictions. Examples were extracted from the binary tree diagram shown in Figure 5. AF2 3D structure models, colored according to the pLDDT values, are shown, along with the positions of the first and last amino acids of the full-VL soluble-like foldable segments (orange balls). The values of the four features are indicated, along with the HCA scores. HCA plots of sequences of the full-VL soluble-like foldable segments are also shown (orange, dashed boxes). How to read sequences (1D) and secondary structures (2D) is shown in the inset, as well as the special symbols used to designate four amino acids with respect to their particular structural behavior. Regular secondary structures (RSS), as observed in the AF2 3D structure models, are designated with orange numbers, which are also reported below the HCA plot in order to indicate the correspondence with hydrophobic clusters. RSS predicted only according to the presence of hydrophobic clusters are reported in other colors, and their positions are indicated on the AF2 3D structure models (with the first and last amino acids shown in atomic details). The hydrophobic cluster affinities for RSSs, calculated using only the binary pattern information, are indicated, as extracted from HCDB v2 [36]. The upper (H,E) and lower (h,e) cases stand for strong and weak preferences, respectively. H stands for alpha-helix, E for beta-strand.

Nd stands for hydrophobic clusters for which there are insufficient statistics in HCDB for the assignment of RSS affinity. TM stands for Transmembrane. IUPred2 long disorder predictions (DIS) are indicated in orange. Hydrophobic clusters corresponding to two successive regular secondary structures are broken down into their components (vertical lines). The sequence repeat in panel (d) is boxed on the HCA plot, whereas the basic unit of the repeat was extracted from the 3D structures (shown at the left and right ends). The 3D structure at right illustrates the AF2 prediction for a member of the same family as the protein sequence shown on the left. The blue box corresponds to the sequence included in the leaf that is the most populated in the full-VH tree shown in Figure 3. Subfigures a to d correspond to examples with RSS > 0, disorder  $\geq 33.4$  (a) or disorder < 33.4 (b, c, d). Subfigure e corresponds to an example without RSS.

Logically, about five times as many cases are found with a low number of BFD homologs (1274 segments: Figure 5). This reinforces the observation that while assigning a low confidence score, AF2 can propose models even when little evolutionary information is available (Figure 7c,d). A first example (UniProt A0A1I9LP79, Figure 7c) corresponds to an uncharacterized protein from *Arabidopsis thaliana*, whose 3D structure is predicted as a 12-stranded beta-sandwich. A Dali-server search in the PDB database revealed multiple hits with similar structures but with a lower strand content (Z-scores up to 6.6 and sequence identities below 15% (e.g., pdb:4q7g-A, Z-score of 6.6, 8% identity)). Examination of the HCA plot indicated that all the hydrophobic clusters match the regular secondary structures predicted by AF2. This suggests that the basic secondary structure elements are indeed present in the proposed model, arranged correctly or not. However, no conclusion can be drawn in the absence of a sufficient number of homologs (mean BFD homologs per position: 5.08, mean sequence identity > 60%). A second intriguing example is an uncharacterized protein from *Trypanosoma cruzi* (Q4CUB3), consisting of a repeated motif of 70 amino acids (mean BFD homologs per position: 10.16, identity > 80%) (Figure 7d). This is predicted to form a repeated beta-alpha-beta-alpha motif, with the two helices arranged on either side of a central beta sheet of parallel beta-strands, forming an elongated structure with a continuous hydrophobic core. A Dali search revealed structural alignments with different tandem-repeat structures (Z-scores up to 4.4, with sequences identity below 10%), belonging to distinct structural families (armadillo repeats (pdb:6dee-A, Z-score: 4.3, 7% identity), right-handed beta-helix (pdb:5zru-A, Z-score: 4.1, 3% identity; 1bhe-A, Z-score: 4.1, 5% identity), heat repeats (pdb:5loi-A, Z-score: 4.0, 9% identity)). In addition, AF2 predictions made for some homologous sequences correspond to a different repeat fold, always predicted with a low to very low level of accuracy (e.g., Q4CW36\_TRYCC, >80% mean identity on the repeated sequences, AF2 prediction corresponding to a right-handed beta helix, at right on Figure 7d). This suggests that this repeat module may correspond to a novel 3D structure, which deserves to be explored experimentally.

A third example (UniProt Q8WU49, Figure 7a) illustrates a case containing amino acids predicted to be disordered, in contrast to the former. It corresponds to the uncharacterized human protein C7orf33, which is taxonomically restricted to primates (mean BFD homologs per position: 4.49, mean identity 76%). The 3D structure predicted by AF2 corresponds to a beta-sandwich, with seven strands. A Dali search yielded many results with similar structures (Z-scores up to 5.9 and sequence identities below 15% (e.g., pdb:6eon-A, Z-score 5.7, 8% identity)). Examination of the HCA plot indicated that not all the hydrophobic clusters present in the sequence correspond to the regular secondary structures predicted by AF2. Instead, there are at least five hydrophobic clusters that correspond in the AF2 model to large, unstructured coils. Many of these clusters have strong affinity for the extended (beta-strand) state, as deduced from our hydrophobic cluster dictionary [36]. This suggests that the 3D structure of this protein could incorporate these clusters as additional regular secondary structures. Alternatively, as part of this sequence is predicted to be disordered by IUPred2, it is also possible that this sequence corresponds to a disordered compact domain, helping to maintain a metastable/transient interface for target recognition, as discussed for the C-terminal domain of protein 4.1G [49].

A last category of full-VL, long soluble-like foldable segments with poor solvent accessibility are the cases without RSS. Most of these cases correspond to unfolded segments in contact with other, well-folded protein regions under consideration, making them comparable to the principal category described below. However, a few cases correspond to segments that show a tendency to form a hydrophobic core without the presence of true secondary structures (see for instance the case of a protein from *Oryza sativa* in Figure 7e).

These examples indicate that such foldable domains, with very low AF2 pLDDT values but a presence of regular secondary structures interacting with each other, may correspond to original, well-folded structures. These are thus prime targets for experimental investigation, especially in the absence of sufficiently divergent homologous sequences. These include tandem repeats, which are relatively poorly represented in the PDB compared to other folds [50].

### 3.3.2. Full-VL Segments with AF2 Unfolded Models

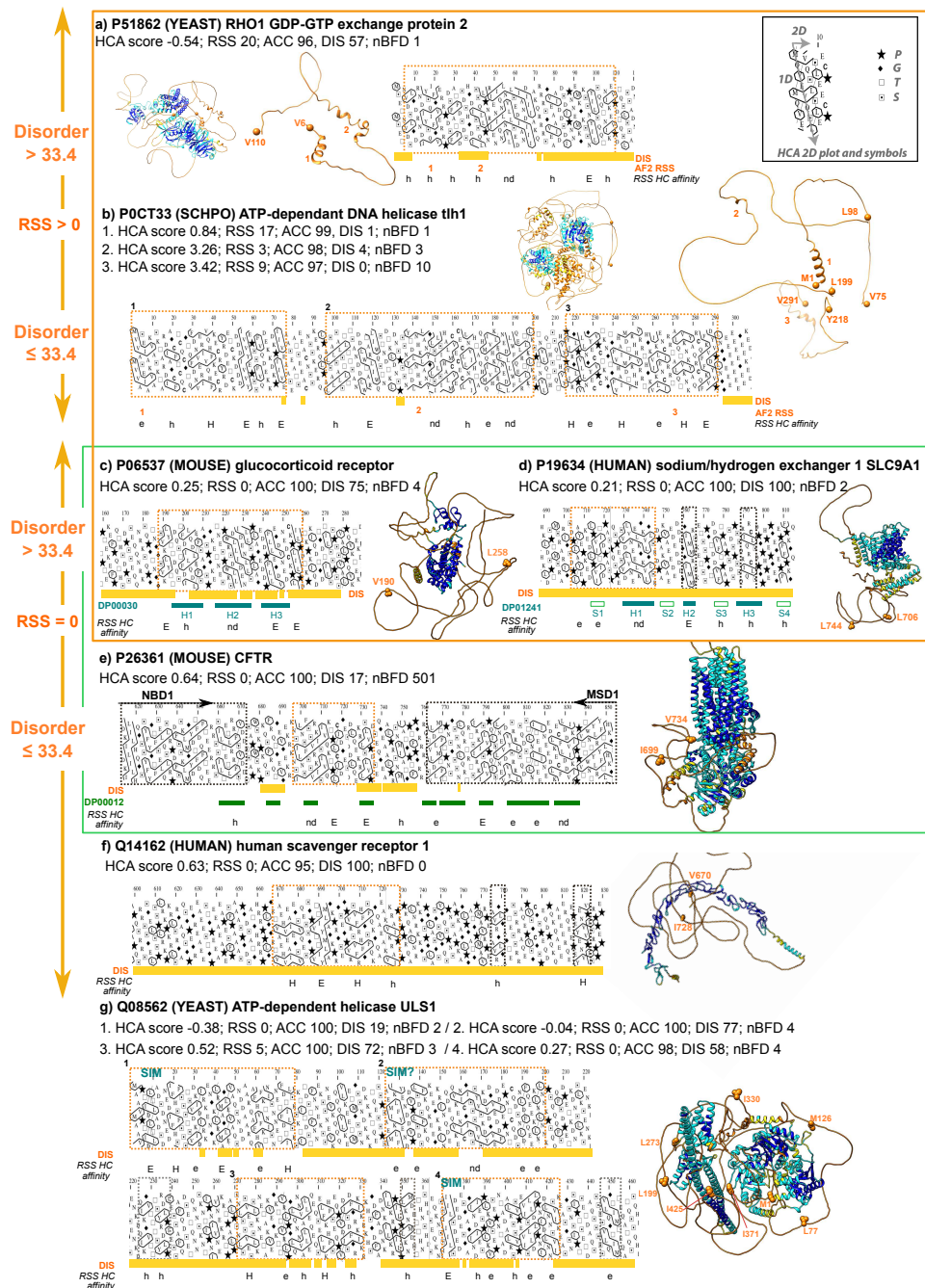
The most abundant category of the full-VL segments (orange boxes in Figure 5) corresponds to unfolded 3D models (encompassing more than 82.9% residues considered solvent accessible by DSSP). These are predicted as disordered or not by IUPred2 and have a low number of known homologs in BFD. This supports the general observation that VL residues are mostly associated with disorder, as no or very few unassembled RSS can be predicted by AF2. The fact that cases with few BFD homologs are about ten times more numerous than cases with a high number of BFD homologs supports the assignment of these segments to the “disorder” category, because IDR sequences are known to be less conserved. However, the HCA score values and the content in hydrophobic clusters suggest that these segments contain conditional order. Nevertheless, it cannot be ruled out that AF2 fails to predict RSS that can assemble into stable, well-folded 3D structures due to the lack of evolutionary information (or, for cases with a high number of BFD homologs, to insufficient depth of multiple sequence alignments). Such cases are referred to as “hidden” (unconditional) order. These hypotheses of conditional or unconditional order cannot be unequivocally demonstrated without the use of experimentation. Nevertheless, we give below some examples supported by experiments that confirm these hypotheses.

The first category (conditional order) is further supported by the fact that some instances are annotated in the DisProt database (Figure 8, green box). This is illustrated by a first example (Figure 8c) corresponding to a foldable segment of the mouse glucocorticoid receptor (GCR, UniProt P06537), including its core transactivation domain (DisProt DP00030, 94.2% identity with human GCR). This domain is intrinsically disordered but forms three helices that are ~30% pre-populated [51]. These three helices correspond to the positions of hydrophobic clusters on the HCA plot.

A second example (Figure 8d) is the foldable segment of the human sodium/hydrogen exchanger 1 (SLC9A1, UniProt P19634), located in its intrinsically disordered intracellular distal tail (aa 686–815, DisProt DP01241). NMR performed on two distant homologs suggested the presence of transient secondary structures and a role in molecular recognition [52]. This role was further supported by a point mutation introduced in the region that disrupts the putative binding feature and impairs trafficking to the plasma membrane [52]. These secondary structures correspond to the positions of hydrophobic clusters on the HCA plots, the first one belonging to the foldable segment described here.

A third example (Figure 8e) is the foldable domain present in the middle of the regulatory (R) region of mouse Cystic Fibrosis Transmembrane conductance Regulator (CFTR, UniProt P26361), a chloride channel belonging to the ABC transporter superfamily (DisProt DP00012, 64% identity with human CFTR). The entire R region of CFTR is a well-known example of an intrinsically disordered sequence whose phosphorylation regulates channel activity [53]. The R region has been shown to interact with the nucleotide-binding domain 1 (NBD1) via multiple transient helices [54]. One of them is included in the foldable region considered here, which is located in the middle of the R region, while the two N- and C-terminal part of the R domain are

embedded in the foldable segments of the preceding (Nucleotide Binding Domain 1) and succeeding (Membrane-Spanning Domain 1) folded domains, respectively.



**Figure 8.** Examples of full-VL soluble-like foldable segments corresponding to unfolded AF2 predictions. See legend in Figure 7. The green box illustrates cases of disordered sequences with transient regular structures (highlighted in green below the HCA plot), documented in the DisProt database (accession number in green at left). SIM stands for Sumo-Interacting Motif. Orange boxes correspond to sequences included in the most populated leaf of the full-VL tree shown in Figure 5. Subfigures a and b correspond to examples with RSS coverage (predicted by AF2) > 0 and disorder coverage (IUPred2 predictions) > 33.4 (a) and ≤ 33.4 (b). Subfigures c to f correspond to examples RSS coverage = 0 and disorder coverage > 33.4 (c, d) and ≤ 33.4 (e, f). Subfigure (g) corresponds to an example with multiple full-VL soluble-like segments, some of which including SLIMs.

Only a small fraction of the foldable segments corresponding to such AF2 predictions (i.e., low pLDDT values, no regular secondary structures, HCA scores typical of folded, soluble domains and high IUPred2 coverage) correspond to sequences included in DisProt, with experimental evidence of conditional disorder. This suggests that the remaining segments, which are numerous, may be interesting targets for experimental studies. One such example is the human scavenger receptor F member 1 (SCARF1 (SREC\_HUMAN), UniProt Q14162, Figure 8f), which plays a key role in the binding and endocytosis of endogenous and exogenous ligand. The importance of SCARF1 in immunological processes was demonstrated using a SCARF1-deficient mice model, which developed systemic lupus erythematosus-like autoimmune disease [55]. A foldable segment (aa 670–728, HCA score: 0.62, IUPred2 coverage 100%) with three hydrophobic clusters typical of an alpha-beta-alpha motif can be found in its large, otherwise, intrinsically disordered cytoplasmic domain (Figure 8d), for which a role in signaling has been suggested but this function has yet to be elucidated [55]. The foldable segment highlighted here is a good candidate for further exploration of conditional order, even though this remains to be supported at the experimental level.

Short linear motifs (SLiMs) [56] are a priori excluded from this study because their lengths are below the threshold fixed here (30 amino acids) and as they often contain only a single hydrophobic cluster [30]. Such cases are associated with higher HCA scores (Bruley et al. [32]). However, some SLiMs can be embedded in larger foldable segments [30], allowing their detection in the present dataset. This is illustrated by four foldable segments detected in the N-terminal region of yeast ULS1 (UniProt Q08562, Figure 8g). This ATP-dependent helicase is required for end-joining inhibition at telomeres and interacting with the silencing regulator Sir4 [57]. SUMO-interacting motifs can be found within the first and fourth foldable segments, while a third can be suspected in the second foldable segment. The advantage of the HCA-based approach is to propose a prediction, through the boundaries of the foldable domains, of the structurally coherent neighborhood of the interacting modules, and thus highlight the sequences that confer flexibility, adaptability, and dynamic character to the IDRs.

Finally, we also observed cases of full-VL, long soluble-like foldable segments with RSS but accessible to solvent (Figure 8a,b). These can be compared to the most populated category without RSS, corresponding to either possible conditional or hidden order. Consideration of disorder predictions can help to distinguish between the different categories.

#### 4. Discussion

It is now widely accepted that the low confidence structural predictions of AF2 correspond mainly to disorder [1,4,11]. In agreement with other investigations [4,5], we have recently shown that a large fraction of these sequences are indeed included in non-foldable segments as defined by pyHCA, which can therefore be considered as “full disorder” [32]. However, a substantial part of sequences with very low confidence scores in AF2 also belongs to foldable segments, in particular, those with a density in hydrophobic clusters typical of soluble domains. This led us to further study their structural characteristics, with respect to the type of order they might contain. The non-foldability/foldability of sequences is estimated by pyHCA from the sole information of a single amino acid sequence, independently of the existence of homologs, whose consideration is one of the pillars of AF2 efficiency.

The key lesson that can be drawn from our study is that the long foldable segments predicted as unfolded by AF2 with very low confidence scores (represented in the form of full-length spaghetti, like those of non-foldable segments), in fact most likely contain either conditional order or hidden, non-conditional order.

Conditional order (or disorder) can be considered as a consequence of the marginal stability of the folded state, making us aware that structure can be determined by both the sequence and the environment [58]. Here, we specifically addressed the issue of intrinsically disordered domains (IDDs), since we only considered long segments (>30 amino acids) that, moreover, are

likely to correspond to homogeneous structural units, according to the definition of foldable segments. Shorter foldable segments, including a large part of MoRFs, belong to another category, characterized by higher HCA scores [32], which was not explored here. It should be noted that short linear motifs (SLiMs) can be embedded in larger foldable segments, constituting the structural unit that can modulate their interaction properties. For instance, the study of CBP interaction domain (CID) of the p160 transcriptional co-activator NCOA3 revealed that its flanking regions promote binding through short-lived, non-specific hydrophobic contacts with the partner [59]. These hydrophobic contacts are provided by hydrophobic clusters that are part of the foldable segments in which CID is included.

A recent study has shown that AF2 predicts 60% of the conditional order with high accuracy, capturing the folded state [5]. This reinforces the assumption that low scoring corresponds to full disorder. Our study provides a refined analysis and new insights for additional conditional order unidentified by AF2, which represent interesting targets worth investigating an experimental level.

The long, soluble-like full-VL foldable segments studied here may correspond to (i) cases of induced folding without the formation of a folded domain, resulting from the interaction of individual regular secondary structures with a partner, (ii) cases where a folded 3D structure is formed, dependent on the partner to be induced/stabilized, (iii) cases where a folded 3D structure is stably formed, independent of the environment (what we designate as unconditional, hidden order). This unconditional order remains completely invisible in AF2 predictions, presumably due to the lack of homologs or insufficient depth of the multiple sequence alignments used in the machine learning process.

While cases of conditional order can be supported by taking into account the DisProt database, this is not obvious for cases of hidden, non-conditional order. These indeed correspond to the unknown part of the proteomes (also described as dark proteomes). However, the HCA characteristics of these foldable segments with unfolded AF2 models (Figure 8) are comparable to those of well-folded AF2 models from the full-VL (Figure 7) and full-VH (Figure 4) categories. This supports the hypothesis that these foldable segments are still unexplored reservoirs of well-folded 3D structures. Whether these sequences correspond to true orphans, or at least taxonomically restricted genes, or whether they share distant relationships that cannot be detected by current homology detection methods is a difficult question to answer. It requires in particular novel methods going beyond sequence similarities. Recent developments for the detection of distant homologs (e.g., [60]) but also for 3D structure prediction from single protein sequences without known homologs (e.g., [61], based on the protein language model) will thus open new perspectives to decipher these cases.

The distinction between conditional and hidden, non-conditional order is not straightforward, but can be guided by taking into account current disorder predictors, in particular integrating more information on the amino acid composition. Useful information could also be given by the hydrophobic cluster composition (e.g., based on the HCA toolkit), as well as by sequences linking the hydrophobic clusters, which correspond mainly to loops.

Several hypotheses can explain the low confidence scores associated with the folded AF2 model segments. First, the proposed 3D structures should be adopted but are not yet validated by AF2 due to either: original folds/structures, the lack of representation in the databases used for learning, or an insufficient amount of homologous sequences to validate the predicted contacts. This hypothesis was recently supported in particular by Sen and colleagues [62], showing lower AF2 pLDDT values for models of sequences corresponding to unassigned domains, compared to those corresponding to CATH or Pfam entries.

Second, the proposed 3D structures should not be adopted, due to incorrect RSS assembly, with sometimes some RSSs not yet well predicted. Nevertheless, the signature of folding is there and thus, given that these proteins are largely uncharacterized, they constitute interesting targets for experimental validation, and characterization of new functions. Among these uncharacterized



sequences are de novo gene candidates, as illustrated with the yeast YBR032W protein in Figure 7b [63]. Other cases are protein repeats, which are widespread periodic units involved in a wide range of functions but are generally difficult to predict due to artifacts resulting from inherent translational symmetry [64]. At the protein level, the structural mechanisms of orphan gene emergence remain to be understood. A fine-grained exploration of foldable segments within the expanding reported cases in eukaryotic proteomes (e.g., *Drosophila* [65], *Oryza* [66], Yeast [63]) would shed light on a still open debate related to the suggested disordered nature of de novo proteins, as a first structural intermediate after gene birth (e.g., [67–71]).

**Supplementary Materials:** The following supporting information can be downloaded at: [www.mdpi.com/xxx/s1](http://www.mdpi.com/xxx/s1), Figure S1: Full-VH binary tree diagrams per proteome; Figure S2: HCA plots of the sequences of full-VH soluble-like foldable segments; Figure S3: Full-VL binary tree diagrams per proteome; Table S1: Distribution of VH and VL residues within the long soluble-like foldable segments of the AFDDB v1 dataset (21 proteomes).

**Author Contributions:** Conceptualization, E.D. and I.C.; methodology, A.B., E.D. and I.C.; software, A.B.; validation, A.B., E.D. and I.C.; formal analysis, A.B., E.D. and I.C.; investigation, A.B., J.-P.M., E.D. and I.C.; resources, A.B.; data curation, A.B., E.D. and I.C.; writing—original draft preparation, E.D., I.C.; writing—review and editing, A.B., J.-P.M., E.D. and I.C.; visualization, A.B., E.D., I.C.; supervision, E.D. and I.C.; project administration, I.C.; funding acquisition, E.D. and I.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** A.B. was supported by the PhD program of Doctoral School “Complexité du Vivant” (ED515, Sorbonne Université). This work was supported by the French National Research Agency (PHOSTORE: ANR-19-CE01-0005 and APOTHESES: ANR-21-CE12-0021).

**Institutional Review Board Statement:** Not applicable

**Informed Consent Statement:** Not applicable

**Data Availability Statement:** The data that support the findings of the present study are available upon request from the corresponding authors E.D. and I.C.

**Acknowledgments:** Analyses were processed with the support of the computer cluster “Plateforme Calcul Intensif Algorithmique” (UMS2700-PCIA) of the Muséum National d’Histoire Naturelle (MNHN).

**Conflicts of interest:** The authors declare no conflicts of interest.

## References

- Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
- Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G.R.; Wang, J.; Cong, Q.; Kinch, L.N.; Schaeffer, R.D.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876. <https://doi.org/10.1126/science.abj8754>.
- Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; et al. AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **2022**, *50*, D439–D444. <https://doi.org/10.1093/nar/gkab1061>.
- Akdel, M.; Pires, D.E.V.; Porta Pardo, E.; Jänes, J.; Zalevsky, A.O.; Mészáros, B.; Bryant, P.; Good, L.L.; Laskowski, R.A.; Pozzati, G.; et al. A structural biology community assessment of AlphaFold 2 applications. *bioRxiv* **2021**. <https://doi.org/10.1101/2021.09.26.461876>.
- Alderson, T.R.; Pritišanac, I.; Moses, A.M.; Forman-Kay, J.D. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *bioRxiv* **2022**. <https://doi.org/10.1101/2022.02.18.481080>.
- Binder, J.L.; Berendzen, J.; Stevens, A.O.; He, Y.; Wang, J.; Dokholyan, N.V.; Oprea, T.I. AlphaFold illuminates half of the dark human proteins. *Curr. Opin. Struct. Biol.* **2022**, *74*, 102372. <https://doi.org/10.1016/j.sbi.2022.102372>.
- Porta-Pardo, E.; Ruiz-Serra, V.; Valentini, S.; Valencia, A. The structural coverage of the human proteome before and after AlphaFold. *PLoS Comput. Biol.* **2022**, *18*, e1009818. <https://doi.org/10.1371/journal.pcbi.1009818>.

8. Ruff, K.M.; Pappu, R.V. AlphaFold and Implications for Intrinsically Disordered Proteins. *J. Mol. Biol.* **2021**, *433*, 167208. <https://doi.org/10.1016/j.jmb.2021.167208>.
9. Tang, Q.-Y.; Ren, W.; Wang, J.; Kaneko, K. The Statistical Trends of Protein Evolution: A Lesson from AlphaFold Database. *bioRxiv* **2022**. <https://doi.org/10.1101/2022.04.07.487447>.
10. Wilson, C.J.; Choy, W.Y.; Karttunen, M. AlphaFold2: A Role for Disordered Protein/Region Prediction? *Int. J. Mol. Sci.* **2022**, *23*, 4591. <https://doi.org/10.3390/ijms23094591>.
11. Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Židek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; et al. Highly accurate protein structure prediction for the human proteome. *Nature* **2021**, *596*, 590–596. <https://doi.org/10.1038/s41586-021-03828-1>.
12. Necci, M.; Piovesan, D.; CAID Predictors; DisProt Curators; Tosatto, S.C.E. Critical assessment of protein intrinsic disorder prediction. *Nat. Methods* **2021**, *18*, 472–481. <https://doi.org/10.1038/s41592-021-01117-3>.
13. van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R.J.; Daughdrill, G.W.; Dunker, A.K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D.T.; et al. Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **2014**, *114*, 6589–6631. <https://doi.org/10.1021/cr400525m>.
14. Morris, O.M.; Torpey, J.H.; Isaacson, R.L. Intrinsically disordered proteins: Modes of binding with emphasis on disordered domains. *Open Biol.* **2021**, *11*, 210222. <https://doi.org/10.1098/rsob.210222>.
15. Wright, P.E.; Dyson, H.J. Linking folding and binding. *Curr. Opin. Struct. Biol.* **2009**, *19*, 31–38. <https://doi.org/10.1016/j.sbi.2008.12.003>.
16. Mohan, A.; Oldfield, C.J.; Radivojac, P.; Vacic, V.; Cortese, M.S.; Dunker, A.K.; Uversky, V.N. Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.* **2006**, *362*, 1043–1059. <https://doi.org/10.1016/j.jmb.2006.07.087>.
17. Yan, J.; Dunker, A.K.; Uversky, V.N.; Kurgan, L. Molecular recognition features (MoRFs) in three domains of life. *Mol. Biosyst.* **2016**, *12*, 697–710. <https://doi.org/10.1039/c5mb00640f>.
18. Oldfield, C.J.; Cheng, Y.; Cortese, M.S.; Romero, P.; Uversky, V.N.; Dunker, A.K. Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* **2005**, *44*, 12454–12470. <https://doi.org/10.1021/bi050736e>.
19. Csizsmók, V.; Bokor, M.; Bánki, P.; Klement, E.; Medzihradszky, K.F.; Friedrich, P.; Tompa, K.; Tompa, P. Primary contact sites in intrinsically unstructured proteins: The case of calpastatin and microtubule-associated protein 2. *Biochemistry* **2005**, *44*, 3955–3964. <https://doi.org/10.1021/bi047817f>.
20. Fuxreiter, M.; Simon, I.; Friedrich, P.; Tompa, P. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.* **2004**, *338*, 1015–1026. <https://doi.org/10.1016/j.jmb.2004.03.017>.
21. Lee, S.H.; Kim, D.H.; Han, J.J.; Cha, E.J.; Lim, J.E.; Cho, Y.J.; Lee, C.; Han, K.H. Understanding pre-structured motifs (PreSMos) in intrinsically unfolded proteins. *Curr. Protein Pept. Sci.* **2012**, *13*, 34–54. <https://doi.org/10.2174/138920312799277974>.
22. Watson, M.; Stott, K. Disordered domains in chromatin-binding proteins. *Essays Biochem.* **2019**, *63*, 147–156. <https://doi.org/10.1042/ebc20180068>.
23. Borgia, A.; Borgia, M.B.; Bugge, K.; Kissling, V.M.; Heidarsson, P.O.; Fernandes, C.B.; Sottini, A.; Soranno, A.; Buholzer, K.J.; Nettels, D.; et al. Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **2018**, *555*, 61–66. <https://doi.org/10.1038/nature25762>.
24. Tompa, P.; Fuxreiter, M. Fuzzy complexes: Polymorphism and structural disorder in protein–protein interactions. *Trends Biochem. Sci.* **2008**, *33*, 2–8. <https://doi.org/10.1016/j.tibs.2007.10.003>.
25. Sharma, R.; Raduly, Z.; Miskei, M.; Fuxreiter, M. Fuzzy complexes: Specific binding without complete folding. *FEBS Lett.* **2015**, *589*, 2533–2542. <https://doi.org/10.1016/j.febslet.2015.07.022>.
26. Davey, N.E.; Van Roey, K.; Weatheritt, R.J.; Toedt, G.; Uyar, B.; Altenberg, B.; Budd, A.; Diella, F.; Dinkel, H.; Gibson, T.J. Attributes of short linear motifs. *Mol. Biosyst.* **2012**, *8*, 268–281. <https://doi.org/10.1039/c1mb05231d>.
27. Tompa, P.; Fuxreiter, M.; Oldfield, C.J.; Simon, I.; Dunker, A.K.; Uversky, V.N. Close encounters of the third kind: Disordered domains and the interactions of proteins. *Bioessays* **2009**, *31*, 328–335. <https://doi.org/10.1002/bies.200800151>.
28. Williams, R.W.; Xue, B.; Uversky, V.N.; Dunker, A.K. Distribution and cluster analysis of predicted intrinsically disordered protein Pfam domains. *Intrinsically Disord. Proteins* **2013**, *1*, e25724. <https://doi.org/10.4161/idp.25724>.
29. Zhou, J.; Oldfield, C.J.; Yan, W.; Shen, B.; Dunker, A.K. Intrinsically disordered domains: Sequence → disorder → function relationships. *Protein Sci.* **2019**, *28*, 1652–1663. <https://doi.org/10.1002/pro.3680>.
30. Bitard-Feildel, T.; Callebaut, I. Exploring the dark foldable proteome by considering hydrophobic amino acids topology. *Sci. Rep.* **2017**, *7*, 41425. <https://doi.org/10.1038/srep41425>.
31. Perdigião, N.; Heinrich, J.; Stolte, C.; Sabir, K.S.; Buckley, M.J.; Tabor, B.; Signal, B.; Gloss, B.S.; Hammang, C.J.; Rost, B.; et al. Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 15898–15903. <https://doi.org/10.1073/pnas.1508380112>.

32. Bruley, A.; Bitard-Feildel, T.; Callebaut, I.; Duprat, E. A sequence-based foldability score combined with AlphaFold2 predictions to disentangle the protein order/disorder continuum. *Proteins* **2022**, *in revision*. <https://doi.org/10.22541/au.165942382.26445116/v1>.
33. Bitard-Feildel, T.; Lamiable, A.; Mornon, J.-P.; Callebaut, I. Order in disorder as observed by the "Hydrophobic Cluster Analysis" of protein sequences. *Proteomics* **2018**, *18*, e1800054. <https://doi.org/10.1002/pmic.201800054>.
34. Callebaut, I.; Labesse, G.; Durand, P.; Poupon, A.; Canard, L.; Chomilier, J.; Henrissat, B.; Mornon, J.-P. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): Current status and perspectives. *Cell Mol. Life Sci.* **1997**, *53*, 621–645. <https://doi.org/10.1007/s000180050082>.
35. Eudes, R.; Le Tuan, K.; Delettré, J.; Mornon, J.-P.; Callebaut, I. A generalized analysis of hydrophobic and loop clusters within globular protein sequences. *BMC Struct. Biol.* **2007**, *7*, 2. <https://doi.org/10.1186/1472-6807-7-2>.
36. Lamiable, A.; Bitard-Feildel, T.; Rebehmed, J.; Quintus, F.; Schoentgen, F.; Mornon, J.P.; Callebaut, I. A topology-based investigation of protein interaction sites using Hydrophobic Cluster Analysis. *Biochimie* **2019**, *167*, 68–80. <https://doi.org/10.1016/j.biochi.2019.09.009>.
37. Faure, G.; Callebaut, I. Comprehensive repertoire of foldable regions within whole genomes. *PLOS Comput. Biol.* **2013**, *9*, e1003280. <https://doi.org/10.1371/journal.pcbi.1003280>.
38. Linding, R.; Russell, R.B.; Neduva, V.; Gibson, T.J. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* **2003**, *31*, 3701–3708. <https://doi.org/10.1093/nar/gkg519>.
39. Mészáros, B.; Erdos, G.; Dosztányi, Z. IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **2018**, *46*, W329–W337. <https://doi.org/10.1093/nar/gky384>.
40. Eddy, S. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **2011**, *7*, e1002195.
41. The UniProt Consortium. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–d489. <https://doi.org/10.1093/nar/gkaa1100>.
42. Steinegger, M.; Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **2018**, *9*, 2542. <https://doi.org/10.1038/s41467-018-04964-5>.
43. Steinegger, M.; Mirdita, M.; Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manifold. *Nat. Methods* **2019**, *16*, 603–606. <https://doi.org/10.1038/s41592-019-0437-4>.
44. Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637. <https://doi.org/10.1002/bip.360221211>.
45. Rost, B.; Sander, C. Conservation and prediction of solvent accessibility in protein families. *Proteins* **1994**, *20*, 216–226. <https://doi.org/10.1002/prot.340200303>.
46. Holm, L. Dali server: Structural unification of protein families. *Nucleic Acids Res.* **2022**, *50*, W210–W215. <https://doi.org/10.1093/nar/gkac387>.
47. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612. <https://doi.org/10.1002/jcc.20084>.
48. Carlson, C.B.; Bernstein, D.A.; Annis, D.S.; Misenheimer, T.M.; Hannah, B.L.; Mosher, D.F.; Keck, J.L. Structure of the calcium-rich signature domain of human thrombospondin-2. *Nat. Struct. Mol. Biol.* **2005**, *12*, 910–914. <https://doi.org/10.1038/nsmb997>.
49. Wang, D.; Wu, S.; Wang, D.; Song, X.; Yang, M.; Zhang, W.; Huang, S.; Weng, J.; Liu, Z.; Wang, W. The importance of the compact disordered state in the fuzzy interactions between intrinsically disordered proteins. *Chem. Sci.* **2022**, *13*, 2363–2377. <https://doi.org/10.1039/d1sc06825c>.
50. Kajava, A.V. Tandem repeats in proteins: From sequence to structure. *J. Struct. Biol.* **2012**, *179*, 279–288. <https://doi.org/10.1016/j.jsb.2011.08.009>.
51. Kim, D.-H.; Wright, A.; Han, K.-H. An NMR study on the intrinsically disordered core transactivation domain of human glucocorticoid receptor. *BMB Rep.* **2017**, *50*, 522–527. <https://doi.org/10.5483/bmbrep.2017.50.10.152>.
52. Nørholm, A.B.; Hendus-Altenburger, R.; Bjerre, G.; Kjaergaard, M.; Pedersen, S.F.; Kragelund, B.B. The intracellular distal tail of the Na<sup>+</sup>/H<sup>+</sup> exchanger NHE1 is intrinsically disordered: Implications for NHE1 trafficking. *Biochemistry* **2011**, *50*, 3469–3480. <https://doi.org/10.1021/bi1019989>.
53. Ostedgaard, L.S.; Baldursson, O.; Vermeer, D.W.; Welsh, M.J.; Robertson, A.D. A functional R domain from cystic fibrosis transmembrane conductance regulator is predominantly unstructured in solution. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 5657–5662. <https://doi.org/10.1073/pnas.100588797>.
54. Baker, J.M.R.; Hudson, R.P.; Kanelis, V.; Choy, W.-Y.; Thibodeau, P.H.; Thomas, P.J.; Forman-Kay, J.D. CFTR regulatory region interacts with NBD1 predominantly via multiple transient helices. *Nat. Struct. Mol. Biol.* **2007**, *14*, 738–745. <https://doi.org/10.1038/nsmb1278>.
55. Patten, D.A. SCARF1: A multifaceted, yet largely understudied, scavenger receptor. *Inflamm. Res.* **2018**, *67*, 627–632. <https://doi.org/10.1007/s00011-018-1154-7>.

56. Weatheritt, R.J.; Luck, K.; Petsalaki, E.; Davey, N.E.; Gibson, T.J. The identification of short linear motif-mediated interfaces within the human interactome. *Bioinformatics* **2012**, *28*, 976–982. <https://doi.org/10.1093/bioinformatics/bts072>.
57. Lescasse, R.; Pobięga, S.; Callebaut, I.; Marcand, S. End-joining inhibition at telomeres requires the translocase and polySUMO-dependent ubiquitin ligase Uls1. *EMBO J.* **2013**, *32*, 805–815. <https://doi.org/10.1038/emboj.2013.24>.
58. Hausrath, A.C.; Kingston, R.L. Conditionally disordered proteins: Bringing the environment back into the fold. *Cell Mol. Life Sci.* **2017**, *74*, 3149–3162. <https://doi.org/10.1007/s00018-017-2558-1>.
59. Karlsson, E.; Schnatwinkel, J.; Paissoni, C.; Andersson, E.; Herrmann, C.; Camilloni, C.; Jemth, P. Disordered Regions Flanking the Binding Interface Modulate Affinity between CBP and NCOA. *J. Mol. Biol.* **2022**, *434*, 167643. <https://doi.org/10.1016/j.jmb.2022.167643>.
60. Schütze, K.; Heinzinger, M.; Steinegger, M.; Rost, B. Nearest neighbor search on embeddings rapidly identifies distant protein relations. *bioRxiv* **2022**. <https://doi.org/10.1101/2022.09.04.506527>.
61. Chowdhury, R.; Bouatta, N.; Biswas, S.; Rochereau, C.; Church, G.M.; Sorger, P.K.; AlQuraishi, M. Single-sequence protein structure prediction using language models from deep learning. *bioRxiv* **2021**. <https://doi.org/10.1101/2021.08.02.454840>.
62. Sen, N.; Anishchenko, I.; Bordin, N.; Sillitoe, I.; Velankar, S.; Baker, D.; Orengo, C. Characterizing and explaining the impact of disease-associated mutations in proteins without known structures or structural homologs. *Brief. Bioinform.* **2022**, *23*, bbac187. <https://doi.org/10.1093/bib/bbac187>.
63. Vakirlis, N.; Hebert, A.S.; Opulente, D.A.; Achaz, G.; Hittinger, C.T.; Fischer, G.; Coon, J.J.; Lafontaine, I. A Molecular Portrait of De Novo Genes in Yeasts. *Mol. Biol. Evol.* **2017**, *35*, 631–645. <https://doi.org/10.1093/molbev/msx315>.
64. Espada, R.; Parra, R.G.; Mora, T.; Walczak, A.M.; Ferreiro, D.U. Capturing coevolutionary signals in repeat proteins. *BMC Bioinform.* **2015**, *16*, 207. <https://doi.org/10.1186/s12859-015-0648-3>.
65. Heames, B.; Schmitz, J.; Bornberg-Bauer, E. A Continuum of Evolving De Novo Genes Drives Protein-Coding Novelty in *Drosophila*. *J. Mol. Evol.* **2020**, *88*, 382–398. <https://doi.org/10.1007/s00239-020-09939-z>.
66. Zhang, T.; Faraggi, E.; Li, Z.; Zhou, Y. Intrinsically semi-disordered state and its role in induced folding and protein aggregation. *Cell Biochem. Biophys.* **2013**, *67*, 1193–1205. <https://doi.org/10.1007/s12013-013-9638-0>.
67. Carvunis, A.R.; Rolland, T.; Wapinski, I.; Calderwood, M.A.; Yildirim, M.A.; Simonis, N.; Charleatoux, B.; Hidalgo, C.A.; Barbette, J.; Santhanam, B.; et al. Proto-genes and de novo gene birth. *Nature* **2012**, *487*, 370–374. <https://doi.org/10.1038/nature11184>.
68. Vakirlis, N.; Acar, O.; Hsu, B.; Castilho Coelho, N.; Van Oss, S.B.; Wacholder, A.; Medetgul-Ernar, K.; Bowman, R.W., 2nd; Hines, C.P.; Iannotta, J.; et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat. Commun.* **2020**, *11*, 781. <https://doi.org/10.1038/s41467-020-14500-z>.
69. Wilson, B.A.; Foy, S.G.; Neme, R.; Masel, J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat. Ecol. Evol.* **2017**, *1*, 0146. <https://doi.org/10.1038/s41559-017-0146>.
70. Bitard-Feildel, T.; Heberlein, M.; Bornberg-Bauer, E.; Callebaut, I. Detection of orphan domains in *Drosophila* using "hydrophobic cluster analysis". *Biochimie* **2015**, *119*, 244–253. <https://doi.org/10.1016/j.biochi.2015.02.019>.
71. Bungard, D.; Copple, J.S.; Yan, J.; Chhun, J.J.; Kumirov, V.K.; Foy, S.G.; Masel, J.; Wysocki, V.H.; Cordes, M.H.J. Foldability of a Natural De Novo Evolved Protein. *Structure* **2017**, *25*, 1687–1696.e1684. <https://doi.org/10.1016/j.str.2017.09.006>.





## **Chapitre 4. DISCUSSION**

#### 1.4. Stratégie et objectifs de la thèse :

Au cours de cette thèse, j'ai enrichi l'outil pyHCA, qui permet de délimiter les segments foldables au sein des séquences, avec le score HCA, une métrique permettant de calculer un score de « foldabilité » et d'ainsi positionner une protéine sur un continuum ordre / désordre, en utilisant la seule information de sa séquence en acides aminés. Cette métrique permet de préciser les caractéristiques structurales de séquences qui ne sont pas ou que partiellement annotées par les méthodes usuelles, principalement en raison de l'absence de séquences homologues dans les banques de données. Dans le chapitre précédent nous avons montré comment le score HCA aide à identifier de l'ordre caché dans l'inconnu structural d'AlphaFold2 (prédictions de très faible confiance).

pyHCA peut également être utilisé pour préciser les propriétés structurales de l'inconnu des protéomes (ou *dark proteome*), c'est-à-dire les séquences qui ne sont pas annotées fonctionnellement, ainsi que les séquences pour lesquelles aucun homologue ne peut être identifié dans les banques de données. Comme décrit dans le chapitre 1.3.2, cette absence d'homologues peut résulter de plusieurs causes : (i) il y a des homologues dans les banques de données mais les méthodes de recherche de similarité échouent à détecter de l'homologie lointaine, ou une relation évolutive entre séquences désordonnées, ou (ii) il n'y a pas d'homologues dans les banques de données car la famille de cette protéine n'a jamais été identifiée, qu'elle soit taxonomiquement étendue ou restreinte à un taxon, auquel cas il pourrait s'agir d'un gène ayant récemment émergé à partir de régions précédemment non codantes (gènes *de novo*).

Dans ce chapitre, je souhaite discuter d'un travail exploratoire que j'ai entamé sur les séquences inconnues de deux taxa bactériens (l'espèce *Escherichia coli*, bien décrite et pour laquelle on dispose d'un nombre important de données génomiques et le phylum des cyanobactéries, qui se distingue par sa diversité et présente un sujet d'intérêt au laboratoire). Pour cela, j'ai mis en place une stratégie d'annotation des protéomes combinant classification non supervisée, recherche d'homologie lointaine, et recherche de similitude dans différentes banques de données (méta)génomiques, de manière à limiter autant que possible les biais méthodologiques et d'échantillonnage. A l'issue de cette procédure, les protéomes sont découpés en régions classées dans 3 catégories : (i) sans homologues connus (singletons), (ii) sans fonction connue mais avec homologues connus, et (iii) avec fonction connue. J'ai ainsi pu ensuite utiliser l'outil pyHCA pour décrire spécifiquement chacune de ces trois catégories. Une des questions à laquelle nous avons notamment essayé de répondre est de déterminer si cet inconnu est enrichi en désordre, comme cela a pu être montré par le passé (cf 1.3.2) (Bitard-Feildel et Callebaut, 2017; Perdigo et al., 2015). Je me suis également attachée à quantifier la part de séquences inconnues qui présente des caractéristiques typiques de domaines globulaires solubles, pour y identifier des protéines ou domaines protéiques présentant des caractéristiques structurales proches de celles déjà caractérisées mais dont la fonction n'a pas encore été décrite.

J'ai également appliqué cette analyse aux séquences de protéines issues de gènes *de novo* présentées dans la littérature et dont l'expression a été validée expérimentalement. Pour cette analyse, je me suis concentrée sur deux espèces d'eucaryotes. i) d'une part, l'eucaryote unicellulaire *Saccharomyces cerevisiae*, qui présente une grande plasticité de génome, lui



permettant une évolution et donc une adaptation rapide (Sipiczki, 2011), et dont les gènes *de novo* ont été décrits par plusieurs études (Bungard et al., 2017; Carvunis et al., 2012; Lu et al., 2017; Scaiewicz et Levitt, 2015; Vakirlis et al., 2020; Wu et Knudson, 2018) ; ii) d'autre part, *Oryza sativa*, une céréale au génome complexe et qui présente de forts taux de réarrangements (Tongwu Zhang et al., 2012). Cette analyse me permettra de prendre part au débat relatif aux caractéristiques structurales des protéines ayant émergé *de novo*, présenté en 1.3.2. En effet, deux points de vue s'affrontent, l'un affirmant que les protéines ayant émergé *de novo* présentent un taux de désordre particulièrement élevé, les rendant moins dangereuses pour la cellule car réduisant leur potentiel d'agrégation (Bitard-Feildel et al., 2015; Ekman et Elofsson, 2010; Schmitz et al., 2018; Wilson et Masel, 2011), l'autre affirmant que ces protéines ont des caractéristiques globulaires typiques, mais qu'elles pourraient gagner en désordre avec le temps (Carvunis et al., 2012; Vakirlis et al., 2020, 2018). L'analyse de ces séquences via notre outil devrait nous permettre de les situer sur le continuum désordre-ordre et d'apporter un éclairage nouveau à ce débat. Enfin, j'ai souhaité étudier le lien entre scores HCA des segments foldables présents dans ces protéines *de novo* et scores pLDDT (scores de confiance associé à la prédiction d'AF2), et de les comparer à ceux des segments foldables de l'ensemble des séquences de *S. cerevisiae* et *O. sativa*. Cette étude devrait permettre de préciser quels types de caractéristiques structurales sont présentes dans ces séquences.

Les travaux présentés ci-après sont préliminaires et feront l'objet d'une publication dans le futur.

## 4.1. Identification des séquences de l'inconnu

Pour cette étude, j'ai collecté un jeu de données contenant les séquences de deux taxa bactériens. J'ai choisi l'espèce *Escherichia coli*, bien décrite dans la littérature et facile à cultiver. C'est l'espèce pour laquelle on retrouve la plus grande quantité de génomes complets. On s'attend, d'une part, à retrouver très peu d'inconnu dans son génome et, d'autre part, il serait relativement aisé de caractériser expérimentalement les séquences inconnues présentant des caractéristiques structurales d'intérêt. L'autre taxon est le phylum des cyanobactéries, phylum très diversifié en termes d'écologie, physiologie et morphologie. Les cyanobactéries sont un sujet d'intérêt au laboratoire, on dispose d'ailleurs d'une souchothèque de cyanobactéries et d'une expertise dans leur culture.

Pour constituer ce jeu de données, j'ai récupéré les 1056 protéomes complets d'*Escherichia coli*, et les 1093 protéomes de cyanobactéries (complets et partiels, de manière à composer deux jeux de données de taille comparable) disponibles sur le site du NCBI<sup>7</sup> le 6 Mars 2020. Depuis cette date, 1405 protéomes complets d'*E. coli* et 2686 protéomes de cyanobactéries ont été ajoutés (effectifs relevés le 13 octobre 2022), il serait donc important d'actualiser par la suite nos jeux de données pour tenir compte de cet accroissement.

---

<sup>7</sup> <https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/>

#### 4.1. Identification des séquences de l'inconnu

La stratégie que j'ai mis en place pour identifier l'inconnu au sein de ces protéomes permet une classification à l'échelle des régions et non des protéines entières, comme c'est le cas dans Agnostos-DB (Vanni et al., 2021) ou Novel Gene Fam (Río et al., 2022). Cela permettra notamment une analyse fine des caractéristiques de chaque région. Cette stratégie se découpe en 5 étapes, illustrées en Figure 4.1.

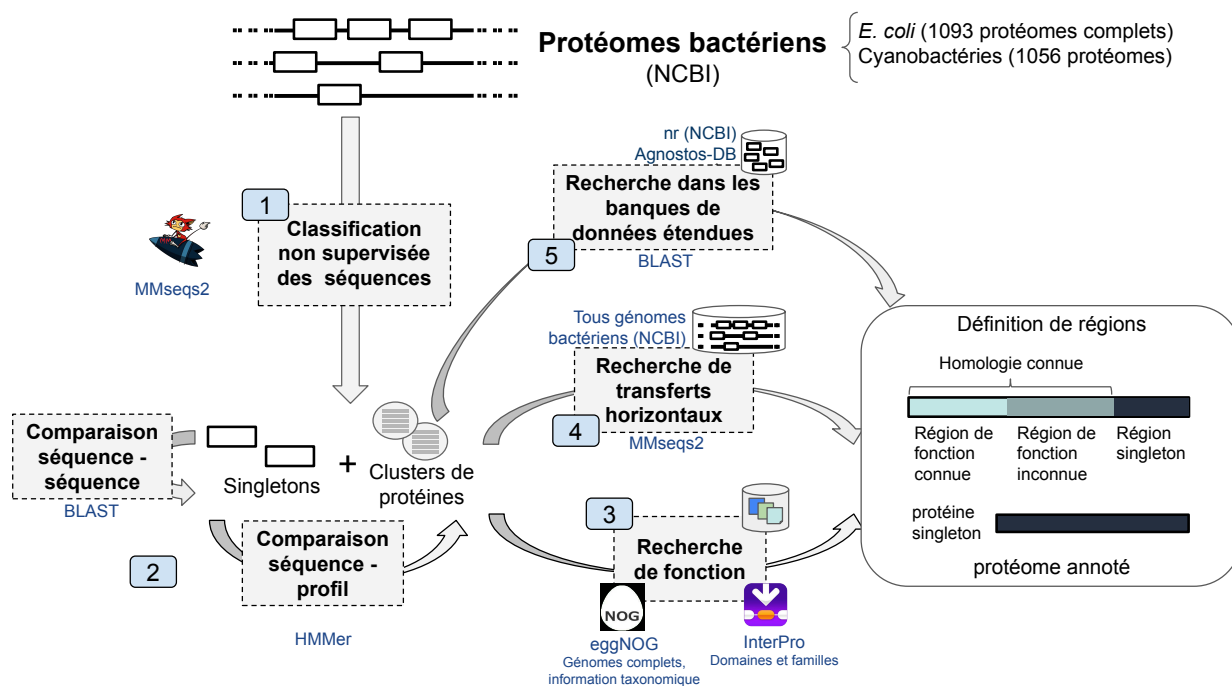


Figure 4.1. Stratégie d'annotation des génomes bactériens. La stratégie est appliquée indépendamment à deux jeux de données : protéomes complets d'*E.coli* et protéomes de cyanobactéries. Les étapes de classification non supervisée (1 et 4) sont réalisées par MMseqs2 avec les paramètres : mode de regroupement = 0, couverture réciproque >80%. Pour les étapes de recherche d'homologues (2,3 et 5) un seuil de e-valeur de  $10^{-4}$  est appliqué.

**Recherche de similitude interne au jeu de données.** J'ai utilisé MMseqs2 (Steinegger et Söding, 2018) pour regrouper en familles les protéines d'un même jeu de données (**étape 1**). A l'issue de cette étape, chaque cluster est considéré comme une famille de protéines, et les protéines laissées en dehors des clusters comme des singletons. Puis j'ai recherché de l'homologie entre clusters et séquences (**étape 2**). J'ai généré les profils HMM à partir de chaque cluster MMseqs2 avec hmmbuild, pour les aligner sur les séquences singletons avec hmmscan. En parallèle, j'ai effectué un alignement par paire des séquences singletons en utilisant blastp (Camacho et al., 2009). Les séquences sans homologie à cette étape restent singletons.

**Recherche de similitude à l'extérieur du jeu de données.** Les séquences de chaque jeu de données ont également été recherchées dans différentes banques de données. Une recherche des séquences représentatives de chaque cluster et toutes les séquences singletons, dans les banques de données eggNOG (Huerta-Cepas et al., 2019, 2017) et d'InterPro (Blum et al., 2021; Jones et al., 2014) permet leur annotation fonctionnelle (**étape 3**). Les régions décrites par des mots-

clés tels que « unknown function » ou « DUF » à l'issue de cette étape seront considérées par la suite comme des régions d'homologie connue mais de fonction inconnue. Pour identifier les singletons qui sont issus de transferts horizontaux j'ai réalisé une seconde classification non supervisée des séquences avec MMseqs2 sur l'intégralité des protéomes bactériens complets, également récupérés le 6 Mars 2020. Seules les séquences qui ne sont pas retrouvées en clusters restent singletons (**étape 4**). Enfin, j'ai réalisé des recherches de similitude avec blastp dans la banque non redondante (nr) du NCBI (Sayers et al., 2021) et parmi les séquences représentatives d'Agnostos-DB (Vanni et al., 2021), qui intègre plusieurs banques de données métagénomiques (Acinas et al., 2021; Kopf et al., 2015; Parks et al., 2018; Yooseph et al., 2007) (**étape 5**). L'utilisation de ces données métagénomiques permet de limiter autant que possible le biais d'échantillonnage des banques de données de séquences.

**Découpage des protéines en régions.** A l'issue de ce processus, j'ai récupéré les bornes des alignements, me permettant de définir les régions au sein des séquences pour lesquelles on retrouve des similitudes avec des séquences dont on connaît la fonction (3), ou non (1,2,4,5 et 3, dans le cas des DUF). Ainsi, par élimination, j'ai pu délimiter des régions singletons. On définit ainsi 3 types de régions : régions de fonction connue, régions avec homologues connus mais de fonction inconnue et régions singletons. Parmi les singletons, on distingue les protéines entièrement singletons des régions singletons, qui flanquent des régions qui peuvent être annotées vis-à-vis de banques de données.

Les effectifs de régions pour chaque catégorie ainsi que le nombre de protéines entièrement singletons sont représentées en Figure 4.2. On observe une proportion importante de régions annotées fonctionnellement dans les deux jeux de données, et en considérant l'intégralité des protéines (44% pour *E. coli* et 34% pour les cyanobactéries), et des régions avec homologie connue (56% et 63% pour *E. coli* et pour les cyanobactéries respectivement). Cependant, lorsqu'on se place à l'échelle des familles, en ne considérant qu'un seul membre de chaque famille (définies à l'étape 1 de la stratégie d'annotation), la proportion de régions connues (en termes de fonction ou d'homologie) diminue et la part de régions et protéines singletons augmente (40% et 37% de régions singletons et 2% et 14% de protéines singletons pour les protéomes d'*E. coli* et de cyanobactéries respectivement). Les protéines annotées fonctionnellement sont donc beaucoup moins diversifiées que les protéines sans annotation fonctionnelles. Cela s'observe pour les deux taxons (on passe de 44% de régions annotées fonctionnellement à l'échelle des protéines à 27% à l'échelle des familles dans les protéomes d'*E. coli* et de 34% à 16 dans les protéomes de cyanobactéries). Cette tendance a été mise en avant par Miller et al., (2022), qui a annoté la fonction des séquences des banques WGS du NCBI et Mgnfy du EBI avec un algorithme d'apprentissage basé sur des algorithmes de traitement automatiques des langues (ou *natural language processing*). Dans ces banques de données, la part d'inconnu passait de 26% lorsque quantifié à l'échelle des familles à 80% lorsque quantifié à l'échelle des protéines.

#### 4.2. Utilisation du score HCA pour décrire de l'inconnu

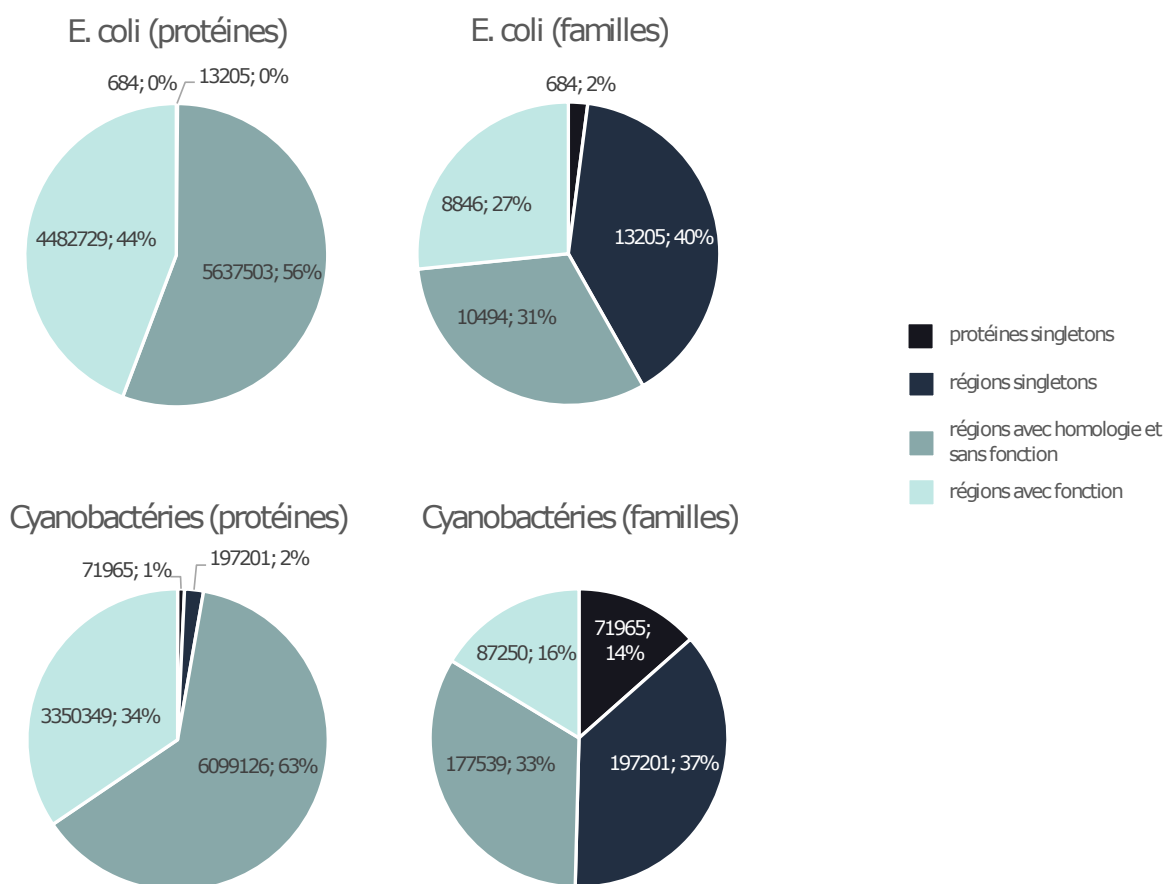


Figure 4.2. Effectif et pourcentage par catégories de connu/inconnu dans les protéomes d'*E. coli* (en haut) et de cyanobactéries (en bas). Les régions ont été comptabilisées sur l'intégralité des protéines de chaque protéome (à gauche) et sur les familles de protéines, définies à l'étape 1 de la stratégie d'annotation (Figure 4.1).

#### 4.2. Utilisation du score HCA pour décrire de l'inconnu

J'ai appliqué la fonction *segment* de pyHCA à mon jeu de données de séquences d'*E. coli* et de cyanobactéries, me permettant ainsi de délimiter les segments foldables et non foldables de celles-ci. Cette information est recoupée avec le découpage en régions décrit ci-dessus, me permettant de calculer, par protéine, le taux de résidus de ces régions retrouvés dans des segments foldables ou non foldables (Figure 4.3).

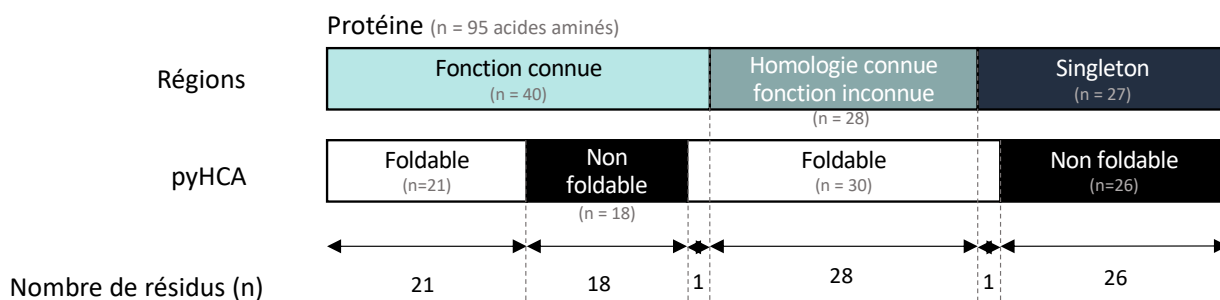


Figure 4.3. Représentation schématique du croisement entre le découpage en régions connues/inconnues (décrit en Figure 4.1) et le découpage en segments foldables par pyHCA. Au sein de cette protéine, on considère ainsi que sont dans des segments foldables : 55% (=22/40) des résidus de fonction connue, 100% (=28/28) des résidus d'homologie connue mais de fonction inconnue et 3.7% (=1/27) des résidus singletons. Les proportions calculées pour l'ensemble des protéines du jeu de données sont représentées en Figure 4.4.

La Figure 4.4 représente le pourcentage, par protéine, d'acides aminés de chaque catégorie (fonction connue, homologie connue mais fonction inconnue et singleton) et présents au sein de segments foldables (cf légende de la Figure 4.3 pour plus de détail). On observe que le pourcentage de résidus en segments foldables diminue avec le niveau d'annotation (médiane pour les régions dont la fonction est annotée : 94.24% et 93.67%, dont la fonction est inconnue mais pour lesquelles on retrouve des homologues : 82.05% et 84.21%, protéines et régions singletons : 66.67% et 76.92% pour les jeux de données respectivement d'*E. coli* et de cyanobactéries). L'inconnu est donc enrichi en désordre, et cette tendance est observée dans les deux jeux de données.

D'autre part, il est intéressant de constater que la part de segments foldables est bien moins importante dans les régions singletons que dans les protéines singletons (médiane pour les régions singletons : 64.29% et 62.5% ; médiane pour les protéines singletons : 88% et 87.88% pour les protéomes respectivement d'*E. coli* et de cyanobactéries). Ce résultat peut s'expliquer par le fait qu'une part importante des régions inconnues retrouvées au sein de protéines partiellement décrites correspond à des boucles ou à des éléments de liaison (« linkers »). Cette hypothèse est largement appuyée par le fait que seulement 6.3% et 1.6% des protéines singletons respectivement d'*E. coli* et de cyanobactéries sont de taille inférieure à 30 acides aminés, contre 84.56% et 72.97% des régions singletons.

Enfin, on peut noter la présence de quatre régions annotées fonctionnellement et sans segments foldables dans les protéomes d'*E. coli* et 319 dans ceux des cyanobactéries. Trois des quatre séquences d'*E. coli* sont situées sur un plasmide et portent l'annotation de protéines impliquées dans la réponse aux stress thermique et acide (Robbe-Saule et al., 2007) et sont désordonnées<sup>8</sup>.

---

<sup>8</sup> <https://www.ebi.ac.uk/interpro/entry/InterPro/IPR019626/>

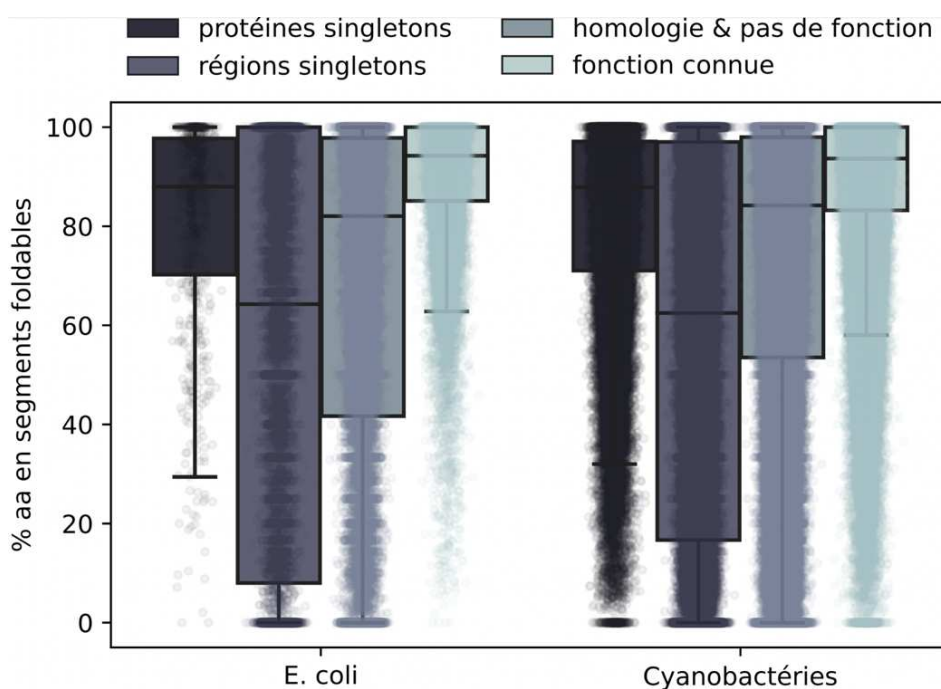


Figure 4.4. Pourcentage de résidus, par protéine, appartenant à une des quatre catégories (protéines singletons, régions singletons, régions avec homologie et fonction inconnue, régions de fonction connue) et retrouvés dans des segment foldables, pour les protéomes d'*Escherichia coli* et de cyanobactéries.

Je me suis ensuite focalisée sur l'étude des segments foldables de chaque catégorie. Je considère qu'un segment foldable appartient à une région donnée s'il est compris à plus de 90% dans celle-ci (illustration en Figure 4.5). D'après la distribution des scores HCA présentée en Figure 4.6., on peut observer que dans les quatre catégories étudiées, très peu voire pas de segments foldables présentent des scores spécifiques du désordre (score HCA < -4.7). Au contraire, la part de segments foldables contenant un seul amas (score HCA > 7.6) est non négligeable, notamment dans les catégories contenant de l'inconnu. Elle est de 57.1% et 46% d'acides aminés pour les segments issus de régions singletons respectivement d'*E. coli* et de cyanobactéries, 17.5% et 19.3% pour les segments d'homologie connue et fonction inconnue et de 9.3% et 10.3% pour les segments foldables dont la fonction est annotée. Ces segments foldables peuvent notamment correspondre à des hélices transmembranaires ou à de courts motifs retrouvés dans des régions désordonnées et incluant des éléments de reconnaissance (SLiMS, MoRFs). Préciser leur nature nécessite des analyses supplémentaires, portant notamment sur la nature des amas présents (voir chapitre Perspectives).

		Protéine (n = 95 acides aminés)				
Régions		Fonction connue (n = 40)		Homologie connue fonction inconnue (n = 28)		Singleton (n = 27)
pyHCA		Foldable (n=21)	Non foldable (n = 18)	Foldable (n = 30)	Non foldable (n=26)	
				> 90% du segment foldable		
		Fonction connue	Non foldable	Homologie connue fonction inconnue	Non foldable	

Figure 4.5. Illustration de la procédure d'attribution des segments foldables à une région connue/inconnue. Par exemple, 28 sur les 30 acides aminés du second segment foldable sont inclus dans la région d'homologie connue mais fonction inconnue, donc  $28/30=93.33\%$  du segment est inclus dans cette région, on peut donc considérer que ce segment foldable appartient à la catégorie « homologie connue mais fonction inconnue » étant donné que le seuil de couverture de 90% est atteint.

On observe une évolution du pourcentage de segments ayant un score typique des domaines globulaires solubles ( $-1 \leq \text{score HCA} \leq 3.5$ ) avec le niveau d'annotation des régions : ce pourcentage augmente quand on passe des régions singletons (27.1 et 33.5% segments foldables respectivement d'*E. coli* et de cyanobactéries), aux régions avec homologie mais sans fonction (53% et 52.8%), puis aux régions de fonction connue (68% et 67.1%). D'autre part, les protéines singletons présentent plus de segments ayant des scores typiques des domaines globulaires solubles que les régions singletons (38.2% et 47.8%). Ces résultats appuient la constatation que l'inconnu est enrichi en désordre. Dans leur étude, Bitard-Feildel & Callebaut (2017) sont arrivés à une conclusion similaire en observant le contenu en désordre (annoté via MobiDB (Piovesan et al., 2020)) des segments foldables des protéines d'UniProt dont la structure ou la fonction est ou non annotée. Ils ont en effet montré qu'en moyenne, les segments foldables des régions et protéines sans annotation ont un contenu moyen de désordre plus important. Malgré un enrichissement de l'inconnu en désordre, on retrouve une quantité non négligeable de segments sans homologues présentant des caractéristiques de domaines globulaires solubles. Dans cette catégorie-là, 82.62% et 85.73% des segments dans les protéines singletons respectivement d'*E. coli* et de cyanobactéries sont longs ( $> 30$  aa) et 48.94% et 68.14% dans les régions singletons d'*E. coli* et de cyanobactéries. Pour la suite, il serait intéressant de prendre en compte la localisation des régions singletons. En effet, la présence de ces régions aux extrémités des protéines pourrait être due à une inexactitude des bornes du CDS prédit, alors qu'une présence en milieu de protéine, en particulier entre deux régions avec fonction et/ou homologie connue permet d'appuyer son existence.

Il est intéressant de constater que les protéomes d'*E. coli* et de cyanobactéries se comportent de la même manière, que ce soit vis-à-vis de leur contenu en segments foldables (Figure 4.4), ou de la distribution des scores HCA des segments foldables (Figure 4.6), et ce malgré les différences séparant ces jeux de données en termes de diversité, de quantité de souches cultivables, de taille des protéomes, ....

#### 4.2. Utilisation du score HCA pour décrire de l'inconnu

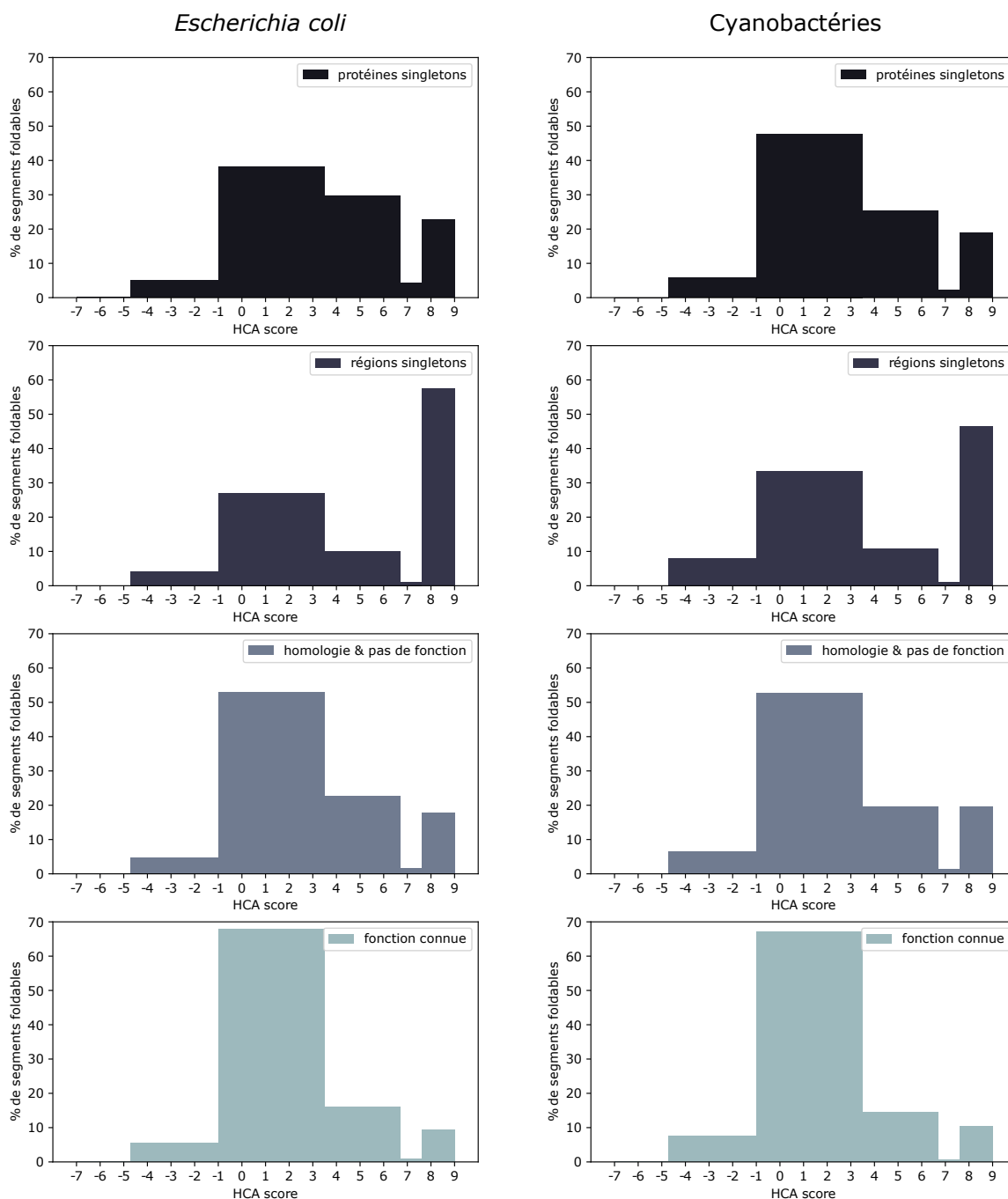


Figure 4.6. Scores HCA des segments foldables compris à 90% ou plus dans les régions singletons, avec homologie mais sans fonction, avec fonction annotée. Les quatre barres couvrent les quatre régions de scores HCA définies au Chapitre 2.

Je me suis ensuite focalisée sur l'étude des protéines singletons contenant les segments foldables ayant des scores typiques de domaines globulaires solubles. C'est parmi ces protéines qu'on retrouvera notamment les gènes *de novo* récents (restreints à une seule souche), des familles de protéines qui ne sont représentées dans aucune des banques de données de séquences ou métagénomiques considérées, ou des séquences qui ont des homologues dans ces banques, mais avec lesquelles on ne détecte pas de similarité (homologie lointaine). De plus, ce signal d'ordre nous indique que ces protéines ont une faible probabilité de correspondre à des artefacts (« spurious », voir chapitre 1.3.2.).



En Figure 4.7 sont représentés deux exemples protéines singletons avec segments foldables typiques de domaines globulaires longs, issus d'un protéome d'*E. coli* (Figure 4.7.a) et d'un protéome de cyanobactérie (Figure 4.7.b). En cohérence avec leur statut de singletons, les prédictions de structure d'AF2 sont majoritairement associées à des scores pLDDT (index de confiance) faibles ou très faibles. Ces deux exemples présentent des tracés HCA denses en amas, typiques de structures de domaines repliés solubles. Cela se reflète dans la composition en structures secondaires de la prédiction d'AF2, bien que tous les amas hydrophobes présents dans la séquence ne soient pas retrouvés sous forme de structures secondaires régulières dans les prédictions (séquence N-terminale de QGY27346.1 et C-terminale de PSB27953.1), ce qui laisse à penser qu'il manquerait peut-être des éléments à la structure prédite par AF2. De plus, ces structures prédites ne sont pas très compactes (avec respectivement 64.4% et 86.81% d'acides aminés accessibles au solvant dans les exemples a et b). Comme nous l'avons discuté en chapitre 2, ces séquences correspondent à des segments susceptibles d'adopter un repliement (de façon non conditionnelle, ainsi que suggéré par leur caractère non désordonné (prédiction IUPred2)). La nature de ce repliement reste encore à préciser dans la mesure où le manque d'information évolutive limite la confiance accordée à la prédiction qui peut être réalisée par AF2.

Même si l'on peut s'attendre à ce que ces protéines contenant des longs segments foldables typiques des domaines globulaires solubles adoptent des repliements déjà répertoriés, il est également possible qu'elles renferment une part de nouveauté. Il serait donc intéressant de les analyser plus avant, afin de sélectionner des candidats à l'expérimentation.

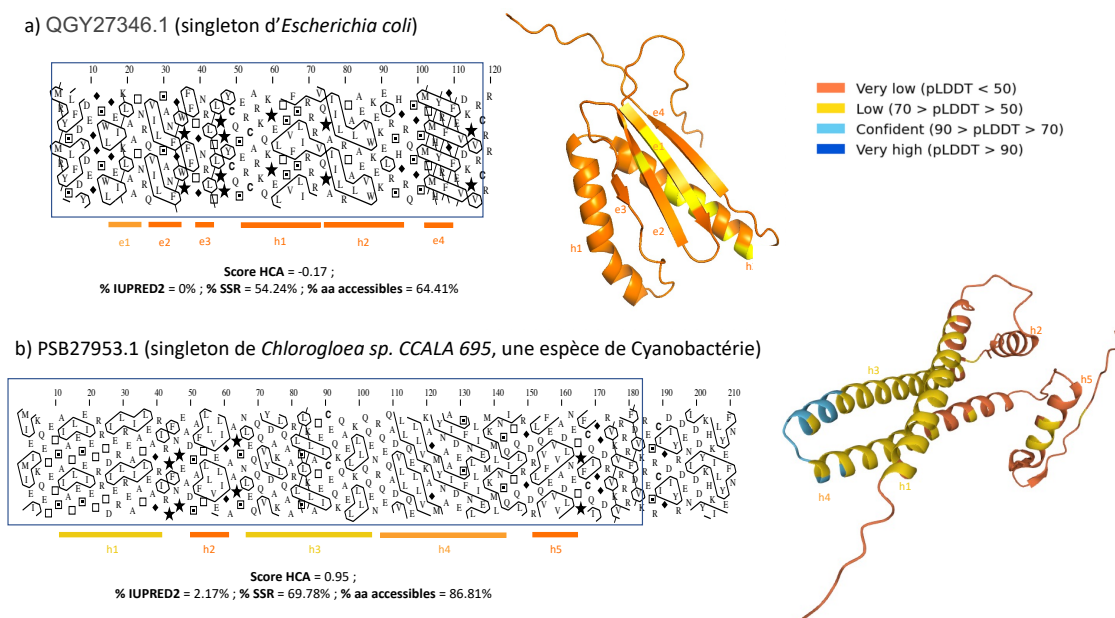


Figure 4.7. Exemple de singletons pour les jeux de données *E. coli* et cyanobactéries. Chaque exemple est illustré par le tracé HCA de la séquence, sur lesquels sont encadrés les segments foldables et par la prédiction de structure 3D d'AlphaFold2. Sous le tracé HCA sont reportées les structures secondaires régulières prédites par AF2. Les segments sont également décrits par leur pourcentage d'acides aminés prédits désordonnés par IUPred2 et par leur pourcentage d'acides aminés inclus dans des structures secondaires régulières et accessibles au solvant dans les modèles de structures 3D d'AF2 (données calculées selon le matériel et méthodes présenté en Chapitre 3).

### 4.3. Utilisation du score HCA pour la description de protéines *de novo*

Les protéines identifiées ci-dessus peuvent correspondre à des cas de *de novo*, elles peuvent également être qualifiées de singletons en raison de biais d'échantillonnage ou de biais méthodologiques. Pour identifier les caractéristiques spécifiques du *de novo*, j'ai donc réalisé ces mêmes analyses sur les protéines ayant émergé *de novo* et spécifiques de *Saccharomyces cerevisiae* et d'*Oryza sativa*.

Dans ce but, j'ai récolté un jeu de données de protéines à partir de gènes *de novo* décrits dans la littérature et validés expérimentalement. Les gènes *de novo* de *S. cerevisiae* ont été extraits du travail présenté par Vakirlis et al. (2018), j'y ai sélectionné les 72 gènes *de novo* qui avaient été détectés par des expériences protéomiques, dont l'expression a donc été vérifiée. Je me suis par ailleurs appuyée sur la publication de Zhang et al. (2019) pour récupérer 175 protéines issues de gènes *de novo* dans le génome d'*O. sativa*. J'ai sélectionné les gènes dont l'expression avait été observée dans au moins une des 10 souches testées (*O. sativa subspecies japonica and indica*, *O. rufipogon*, *O. nivara*, *O. glaberrima*, *O. barthii*, *O. glumaepatula*, *O. meridionalis*, *O. punctata*, *O. brachyantha*).

Ces protéines *de novo* ont été comparées aux protéomes de référence des espèces correspondantes disponibles dans UniProt. Ces protéomes de référence sont des protéomes complets qui sont soit sélectionnés par échantillonnage dans l'arbre du vivant, soit d'intérêt biomédical et biotechnique<sup>9</sup>. Celui de *S. cerevisiae* (identifiant UniProt UP000002311) regroupe 6060 protéines, celui de d'*O. sativa* (UP000059680) contient 48899 protéines. Ces deux protéomes sont représentés dans l'AlphaFold DataBank (Varadi et al., 2022).

La première étape a été d'évaluer l'enrichissement en désordre des protéines *de novo*, comme cela a été discuté dans la littérature. Une première analyse de la composition en segments foldables des protéines *de novo*, par rapport à l'ensemble des protéines des protéomes de référence est représentée en Figure 4.8. On peut observer que dans les deux espèces, la médiane du pourcentage de résidus en segments foldables est plus faible pour les protéines *de novo* (84.61% et 57.42% pour respectivement *S. cerevisiae* et *O. sativa*) que pour les protéomes de référence (87.70% et 79.27%). Il est également intéressant de noter que toutes les séquences *de novo* présentent au moins un segment foldable, aucune n'est complètement désordonnée. Le *de novo* serait donc enrichi en désordre, mais tout en présentant toujours une part d'ordre.

---

<sup>9</sup> [https://www.uniprot.org/help/reference\\_proteome](https://www.uniprot.org/help/reference_proteome)

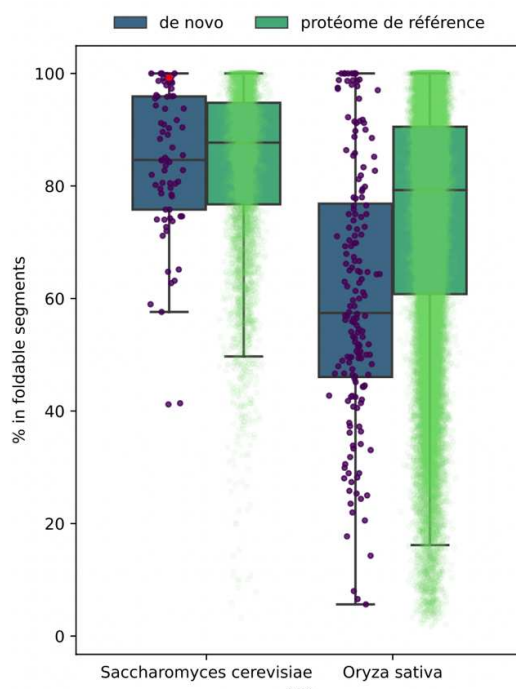


Figure 4.8. Couverture en segments foldables des protéines issues de gènes de novo (en bleu) et de l'ensemble des protéines retrouvées dans chacune des espèces.

Je me suis ensuite focalisée sur les segments foldables de ces jeux de données. *Saccharomyces cerevisiae* et *Oryza sativa* font partie des 21 organismes de référence présents dans la banque de données AFDB. Cela nous permet de recouper les informations apportées par le score HCA aux informations apportées par le score pLDDT d'AlphaFold2 (AF2). J'ai identifié les protéines *de novo* parmi les séquences des protéomes d'AFDB à l'aide d'une recherche blast, en fixant un seuil d'identité de 100% et un seuil de couverture réciproque à 100%. Parmi les 72 séquences protéiques *de novo* de levure, 46 sont retrouvées dans le protéome de référence de levure dont les prédictions 3D sont exploitées dans AFDB. Parmi les 175 séquences protéiques *de novo* du riz, 21 sont dans le protéome de référence d'*O. sativa*. On retrouve 851 segments foldables *de novo* pour *S. cerevisiae* et 499 pour *O. sativa*.

La répartition des scores HCA des segments foldables du *de novo* de *S. cerevisiae* est similaire à celle de son protéome de référence (Figure 4.9.a), les deux distributions présentent notamment un maximum de segments foldables ayant des scores HCA typiques de domaines globulaires solubles. La tendance est la même pour les segments foldables du *de novo* d'*O. sativa* (Figure 4.9.b), qui présente cependant une proportion de segments foldables à un amas plus importante que celle observée pour l'intégralité du protéome.

Dans les protéomes des deux espèces, on observe que les confiances de prédiction d'AF2 sont moins bonnes dans le *de novo*. La différence est notamment marquée pour les segments foldables typiques des domaines globulaires solubles ( $-1 \leq \text{score HCA} \leq 3.5$ ). Les gènes *de novo* présentent par définition moins d'homologues, et aucun homologue distant. AF2 ne pourra donc pas utiliser l'information de coévolution pour inférer les contacts entre résidus au sein de la protéine, ce qui explique la faiblesse des pLDDT des prédictions.

4.3. Utilisation du score HCA pour la description de protéines de novo

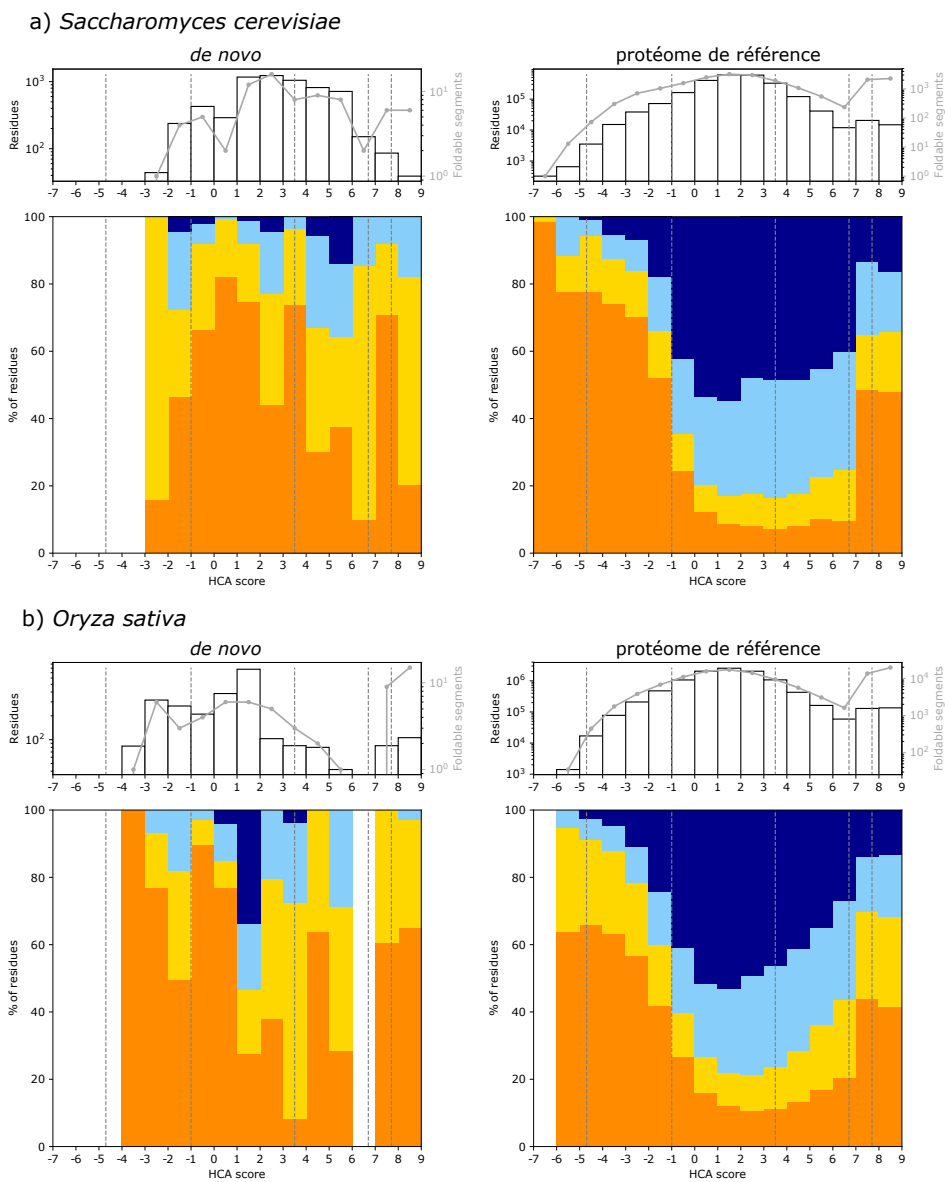


Figure 4.9. Barplots représentant les pourcentages de résidus prédits avec très grande (bleu foncé), grande (bleu clair), faible (jaune) et très faible (orange) confiance, par intervalle de score HCA. Au-dessus de chaque barplot est représenté le nombre de résidus (barres) et de segments foldables (courbe) par intervalle de score HCA. Les pointillés gris délimitent les zones définies dans le chapitre 2.

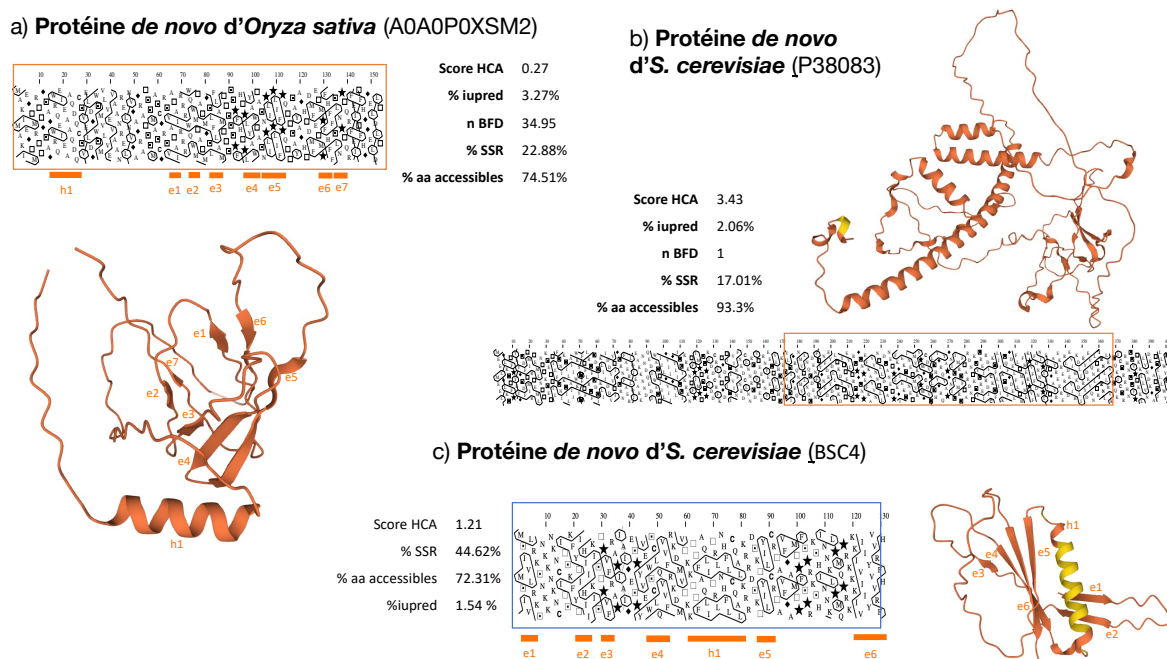


Figure 4.10. Exemples de gène de novo issus des protéomes d'*Oryza sativa* (a) et *Saccharomyces cerevisiae* (b,c).

Parmi les segments foldables longs (> 30 acides aminés) typiques des domaines globulaires solubles, on retrouve 6 segments d'*O. sativa* et 4 de *S. cerevisiae* prédits avec une très faible confiance, et aucun segment foldable long prédit avec très grande confiance. J'ai sélectionné les exemples présentant la plus faible couverture en désordre tel que prédite par IUPred2 (cette couverture varie de 3% à 97% pour *O. sativa* et 2% à 49% pour *S. cerevisiae*), ceux-ci sont présentés en Figure 4.10.a et b. Dans les deux cas, la couverture en structures secondaires régulières des modèles de structure 3D d'AF2 est d'environ 20%, n'incluant qu'une partie des amas hydrophobes présents dans les séquences. Les modèles de structures 3D proposés par AF2 semblent donc incomplets. Cela pourrait être expliqué par le fait que ces séquences présentent des structures voire des repliements originaux, qui mériteraient d'être explorés expérimentalement.

L'exemple illustré Figure 4.10.c correspond à un cas d'un gène *de novo* bien décrit dans la littérature (Bungard et al., 2017). Il a été montré que la protéine BSC4 encodée par ce gène a la capacité de se structurer, en accord avec le segment foldable qui la recouvre dans son intégralité, mais avec des propriétés inhabituelles. AF2 propose une structure peu compacte, avec quelques éléments de structures secondaires régulières séparés par de longues boucles.

En conclusion, cette première analyse de séquences de protéines *de novo* a permis de mettre en avant que celles-ci présentent un enrichissement en désordre, tout en conservant des signatures de repliement. En effet, il n'existe pas de séquences sans segment foldable et les segments foldables présentent des caractéristiques similaires à celles que l'on peut observer pour des domaines globulaires solubles.

Une analyse plus fine de ces protéines pourrait permettre de mettre en avant certaines nuances dans la description de leurs caractéristiques. Comme décrit dans l'étude de Vakirlis et collaborateurs (Vakirlis et al., 2020), le phénomène d'émergence des gènes *de novo* dépend vraisemblablement de la nature de la séquence intergénique. En particulier, les séquences issues de régions intergéniques riches en thymine auraient tendance à coder des domaines transmembranaires putatifs. Ce n'est pas une caractéristique que nous pouvons mettre en avant ici (Figure 4.9, la proportion de segments foldables typiques du membranaire est réduite dans le *de novo*), mais des analyses complémentaires en lien avec la composition des séquences intergéniques seraient nécessaires pour émettre de plus amples conclusions.







## **Chapitre 5. CONCLUSION ET PERSPECTIVES**

Dans le contexte actuel de croissance continue du volume de données de séquences génomiques, décrire automatiquement la fonction des séquences codantes représente un enjeu important. Les progrès récents réalisés pour la prédiction de structures 3D de protéines par des approches telles qu'AlphaFold2 (Jumper et al., 2021) ou RoseTTAFold (Baek et al., 2021) ont conduit à réduire la part d'inconnu structural au sein des protéomes. Comme précisé dans l'introduction, une première étude réalisée par Porta-Pardo et collaborateurs (Porta-Pardo et al., 2022) a estimé que la couverture structurale du protéome humain est passée de 48 % à 76 % d'acides aminés. Ces avancées ouvrent ainsi aujourd'hui de larges perspectives pour améliorer l'annotation fonctionnelle des protéomes, par détection de similitudes entre structures dans les cas où des parentés évolutives éloignées restent difficiles voire impossible à détecter sur la base de la seule comparaison des séquences (Mahlich et al., 2018). De nouvelles méthodes comme Foldseek ou 3D-SURFER, utilisant une description des architectures protéiques à l'aide respectivement d'alphabets structuraux (Kempen et al., 2022) ou de descripteurs de surface moléculaire (3D Zernike descriptors (La et al., 2009)), ont été développées pour la comparaison de structures à haut débit, permettant de pallier aux vitesses limitées des méthodes de détection de similitudes de structures utilisées jusqu'à présent, telles que Dali (Holm, 2020; Holm et Sander, 1996), CE (Shindyalov et Bourne, 1998) ou TM-align (Zhang et Skolnick, 2005). L'utilisation de Foldseek sur les modèles d'AlphaFold2 avec des approches de type RBH (reciprocal best hits) a ainsi conduit à identifier des parentés très éloignées entre séquences (Monzon et al., 2022b). Des approches comme SWORD2 (Cretin et al., 2022) pourraient également être utilisées dans ce contexte pour décomposer les structures en « unités protéiques » et identifier au sein de structures variées des régions fonctionnelles plus conservées d'un point de vue structural.

Si ces avancées permettent des progrès significatifs pour l'annotation des protéomes, il n'en reste pas moins qu'une partie de ceux-ci, non limitée aux séquences intrinsèquement désordonnées, reste inaccessible à ces prédictions structurales. Ainsi, la même étude de Porta-Pardo et collaborateurs (Porta-Pardo et al., 2022) a estimé que cet ordre inconnu représenterait 10 % de l'ensemble des acides aminés du protéome humain. La part d'inconnu d'autres protéomes, par exemple de *Plasmodium falciparum*, est significativement plus élevée, en particulier en raison de la difficulté à identifier des parentés entre ses séquences et celles d'autres organismes, liée à d'importants biais compositionnels (Hamilton et al., 2017; Pizzi et Frontali, 2001).

L'outil que j'ai développé au cours de cette thèse permet d'explorer plus avant les caractéristiques de cet ordre structural inconnu (ou inconnu foldable), inclus dans les prédictions AlphaFold2 de très faible confiance, en positionnant ces segments sur un continuum ordre-désordre (Chapitre 2). Cet ordre encore inconnu contient un grand nombre de régions pour lesquelles les scores de foldabilité, reflétant le ratio ordre/désordre, sont proches de ceux observés pour des domaines globulaires solubles. C'est pour ces mêmes scores de foldabilité qu'on observe un maximum de prédictions AlphaFold2 de (très) bonne qualité, ce qui suggère que cet inconnu foldable soluble est construit à partir de briques semblables à celles que l'on peut observer dans le répertoire des repliements protéiques. Les études que nous avons réalisées pour explorer plus avant les caractéristiques de ces séquences (Chapitre 3) montrent que les

faibles scores de confiance associés aux modèles proposés par AlphaFold2 (que ceux-ci soient repliés ou non) résultent de l'absence d'information évolutive (ou d'une profondeur insuffisante des alignements multiples) et/ou de la présence de structures/repliements inédits, aux caractéristiques non encore répertoriées (et donc absentes de la connaissance sur laquelle se fait l'apprentissage du réseau de neurones d'AF2). Nos résultats donnent accès à ce réservoir d'inconnu foldable soluble, qui constitue un ensemble de candidats privilégiés pour de futures caractérisations expérimentales. Parmi ceux-ci figurent un certain nombre de séquences répétées, pour lesquelles la prédiction de structure 3D constitue toujours un défi de taille, en raison des artéfacts engendrés par la symétrie translationnelle de leurs séquences (Espada et al. 2015). Ces structures/repliements inédits devraient permettre d'accéder à la connaissance de nouvelles fonctions pour caractériser plus avant la diversité du vivant, mais également contribuer à enrichir les connaissances fondamentales utiles pour la prédiction des structures 3D naturelles ou la conception *de novo* de protéines (protein design) (Baker, 2019). Dans ce contexte, les données issues de criblage à haut débit de variants expérimentaux et de leur impact sur la fonction des protéines (approches de « deep mutational scanning ») pourraient être particulièrement utiles (Fowler et Fields, 2014; Ng et Fraternali, 2021).

Le segment foldable tel que défini par l'outil *pyHCA* représente une unité cohérente en termes de densité en amas hydrophobes. Dans le cas des régions intrinsèquement désordonnées, différents cas de figure sont observés pour les segments foldables dont les scores HCA sont similaires à ceux observés pour les domaines globulaires solubles (score compris entre -1 et 3.5). En effet, l'ordre conditionnel qui y est observé correspond à des états de « globules fondus », à des structures repliées stabilisées au sein des complexes oligomériques ou encore à des états désordonnés susceptibles d'acquiescer une structure en contact avec un partenaire (repliement induit). Ces derniers cas incluent des segments contenant des courts motifs de reconnaissance (SLiMs), qui sont cependant souvent associés à de plus hautes valeurs de scores HCA ([7.6,9]). En effet, les SLiMs sont souvent constitués d'amas hydrophobes isolés au sein de régions complètement désordonnées. Leur présence au sein de régions foldables plus larges donne alors une information sur les éléments additionnels susceptibles de moduler les interactions auxquelles ils participent.

Dans le futur, il sera intéressant de développer plus avant l'outil présenté ici pour le calcul de score de foldabilité, en incluant d'autres descripteurs permettant de distinguer parmi les segments foldables les cas d'ordre conditionnel (état désordonné en l'absence d'éléments stabilisateurs) et non conditionnel. En effet, on observe une superposition des distributions de scores HCA des segments foldables propres à des régions désordonnées et à des domaines ordonnés tout au long du continuum décrit par le score HCA, en particulier dans la gamme de scores de type globulaire soluble. L'information du score HCA pourrait être complétée par des informations relatives à la composition en acides aminés des segments foldables. Mes premières tentatives de développement de classifieurs supervisés (de type SVM, ou Support Vector Machine) basés sur la composition en acides aminés présentent des résultats encourageants (données non montrées). Nous envisageons d'intégrer également à ce système de score des informations propres à la composition en amas hydrophobes des segments foldables, telles que la taille et le type d'amas hydrophobes, ainsi que la taille et la composition des séquences inter-

amas. La prise en compte de ces dernières informations a récemment permis de montrer que la plupart des cadres ouverts de lecture (ORFs) inter-géniques de levure contiennent les briques élémentaires des repliements de protéines, permettant ainsi l'émergence de gènes *de novo* (Papadopoulos et al., 2021).

Il serait également intéressant de pouvoir aller plus loin sur la classification de repliements des segments foldables, par étude de la co-occurrence des amas hydrophobes et en intégrant les informations de leurs amas voisins. Ces développements pourraient être réalisés par des approches issues de la linguistique (basés sur le « word embedding » et le traitement par des réseaux de neurones profonds des vecteurs ainsi générés) telles que celles utilisant des pLMs (pre-trained protein language models), qui sont actuellement développées pour réaliser des prédictions de structures 3D à partir de séquences uniques (Chowdhury et al., 2021; Weissenow et al., 2022). A l'aide de tels encodages des séquences, il serait également intéressant de suivre les processus de diversification structurale et fonctionnelle des paralogues (Bershtein et al., 2021). Ces développements pourraient ainsi être appliqués afin de déterminer lesquelles des multiples isoformes possibles de gènes eucaryotes pourraient représenter des produits protéiques foldables (Sommer et al., 2022).

Les approches issues du traitement du langage naturel (NLP pour natural language processing) ont récemment été également utilisées pour prédire les fonctions d'un très large ensemble de gènes microbiens, en fonction de leur contexte génomique (Miller et al., 2022). Ces auteurs constatent ainsi que la plupart des gènes bien caractérisés selon le système d'annotations KEGG sont regroupés dans un nombre relativement restreint de familles très larges (constituant le « core genome »), alors que la plupart de la diversité (80% des familles) n'est pas (ou mal) annotée. Des études menées sur un ensemble de protéines d'archaea indiquent également que les familles de séquences spécifiques de lignées sont très largement constituées de protéines dépourvues d'annotations fonctionnelles, alors même qu'elles pourraient avoir joué un rôle majeur dans leur diversification (Méheust et al., 2022). Les travaux que nous avons entrepris pour annoter les séquences de protéomes de deux taxa (*E. coli* et cyanobactéries) et définir ainsi l'inconnu en termes de fonction et/ou de parentés avec d'autres séquences, conduisent à des conclusions similaires (Figure 4.2, Chapitre 4), appuyant ainsi l'importance de cet inconnu en termes de diversité et de potentielle capacité d'adaptation des écosystèmes. Les premières observations relatives aux scores HCA montrent des tendances similaires entre les taxa étudiés, indiquant : (i) que l'inconnu renferme une part importante de segments foldables, et ii) que la part des segments foldables répondant aux caractéristiques des domaines globulaires solubles décroît avec le niveau de connaissance associé aux séquences, alors que la part de petits segments foldables denses en amas hydrophobes augmente. Cette dernière observation suggère qu'une part importante de l'inconnu foldable correspond à des protéines membranaires (en particulier présente dans la catégorie protéines singletons) et à des courts segments, sièges de transitions désordre-ordre (en particulier présents dans la catégorie régions singletons). Des études plus approfondies de la catégorie « membranaire » devraient être réalisées, à l'image de celles que nous avons finalisées pour la catégorie « soluble ». Enfin, à la suite de ce travail préliminaire, il conviendrait d'actualiser les ensembles construits avec les nouvelles séquences et les nouvelles annotations disponibles. Un travail complémentaire

devra également être entrepris pour écarter les cas de « fausses » séquences (gènes « spurious »), correspondant aux séquences nucléotidiques considérées à tort comme codantes.

Les résultats préliminaires que nous avons obtenus pour l'étude des caractéristiques d'un ensemble restreint de protéines *de novo* permettent également d'alimenter le débat dans ce domaine, en indiquant que ces séquences sont bien enrichies en désordre, mais tout en présentant toujours une part d'ordre. Par ailleurs, la distribution des segments foldables propres à ces séquences le long du continuum décrit par le score HCA est sensiblement similaire à celle des protéomes de référence, suggérant que ces séquences adoptent des caractéristiques structurales semblables à celles de l'ensemble des protéines. Néanmoins et logiquement, en raison du faible nombre ou de l'absence de séquences homologues, la qualité des prédictions de structure 3D réalisées par AF2 sur ces séquences, en particulier pour les segments typiques d'organisations globulaires solubles, est très largement de mauvaise qualité, contrastant avec celles des protéomes de référence. Il sera intéressant dans la suite de ce travail d'étudier les quelques cas de séquences *de novo* pour lesquelles les prédictions AF2 sont de très bonne qualité, et d'enrichir le jeu de données avec l'ensemble des séquences de protéines *de novo* actuellement décrites dans la littérature. L'analyse des segments foldables du *de novo* devrait aussi bénéficier des développements complémentaires de l'outil pyHCA, comme présenté précédemment.

# Bibliographie

- Abet, V., Evans, R., Guibbal, F., Caldarelli, S., Rodriguez, R., 2014. Modular construction of dynamic nucleodendrimers. *Angew. Chem. Int. Ed Engl.* 53, 4862-4866. <https://doi.org/10.1002/anie.201402400>
- Acinas, S.G., Sánchez, P., Salazar, G., Cornejo-Castillo, F.M., Sebastián, M., Logares, R., Royo-Llonch, M., et al., 2021. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Commun. Biol.* 4, 1-15. <https://doi.org/10.1038/s42003-021-02112-2>
- Akdel, M., Pires, D.E.V., Pardo, E.P., Jänes, J., Zalevsky, A.O., Mészáros, B., Bryant, P., et al., 2021. A structural biology community assessment of AlphaFold 2 applications. <https://doi.org/10.1101/2021.09.26.461876>
- Akhila, M.V., Narwani, T.J., Floch, A., Maljković, M., Bisoo, S., Shinada, N.K., Kranjc, A., et al., 2020a. A structural entropy index to analyse local conformations in intrinsically disordered proteins. *J. Struct. Biol.* 210, 107464. <https://doi.org/10.1016/j.jsb.2020.107464>
- Akhila, M.V., Narwani, T.J., Floch, A., Maljković, M., Bisoo, S., Shinada, N.K., Kranjc, A., et al., 2020b. Data set of intrinsically disordered proteins analysed at a local protein conformation level. *Data Brief* 29, 105383. <https://doi.org/10.1016/j.dib.2020.105383>
- Albalat, R., Cañestro, C., 2016. Evolution by gene loss. *Nat. Rev. Genet.* 17, 379-391. <https://doi.org/10.1038/nrg.2016.39>
- Alberti, S., 2017. Phase separation in biology. *Curr. Biol. CB* 27, R1097-R1102. <https://doi.org/10.1016/j.cub.2017.08.069>
- Alderson, T.R., Pritišanac, I., Moses, A.M., Forman-Kay, J.D., 2022. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. <https://doi.org/10.1101/2022.02.18.481080>
- Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., et al., 2021. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* 39, 105-114. <https://doi.org/10.1038/s41587-020-0603-3>
- Altschuh, D., Lesk, A.M., Bloomer, A.C., Klug, A., 1987. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.* 193, 693-707. [https://doi.org/10.1016/0022-2836\(87\)90352-4](https://doi.org/10.1016/0022-2836(87)90352-4)
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402. <https://doi.org/10.1093/nar/25.17.3389>
- Al-Turaiki, I.M., Mathkour, H., Tourir, A., Hammami, S., 2011. Computational Approaches for Gene Prediction: A Comparative Survey, in: Abd Manaf, A., Zeki, A., Zamani, M., Chuprat, S., El-Qawasmeh, E. (Ed.), *Informatics Engineering and Information Science, Communications in Computer and Information Science*. Springer, Berlin, Heidelberg, p. 14-25. [https://doi.org/10.1007/978-3-642-25453-6\\_2](https://doi.org/10.1007/978-3-642-25453-6_2)
- Andreeva, A., Kulesha, E., Gough, J., Murzin, A.G., 2020. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res.* 48, D376-D382. <https://doi.org/10.1093/nar/gkz1064>
- Andrews, S.J., Rothnagel, J.A., 2014. Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* 15, 193-204. <https://doi.org/10.1038/nrg3520>
- Anfinsen, C.B., Haber, E., Sela, M., White, F.H., 1961. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.* 47, 1309-1314. <https://doi.org/10.1073/pnas.47.9.1309>
- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., et al., 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871-876. <https://doi.org/10.1126/science.abj8754>
- Baker, D., 2019. What has de novo protein design taught us about protein folding and biophysics? *Protein Sci. Publ. Protein Soc.* 28, 678-683. <https://doi.org/10.1002/pro.3588>

- Banani, S.F., Lee, H.O., Hyman, A.A., Rosen, M.K., 2017. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* 18, 285-298. <https://doi.org/10.1038/nrm.2017.7>
- Basu, M.K., Poliakov, E., Rogozin, I.B., 2009. Domain mobility in proteins: functional and evolutionary implications. *Brief. Bioinform.* 10, 205-216. <https://doi.org/10.1093/bib/bbn057>
- Bateman, A., Coghill, P., Finn, R.D., 2010. DUFs: families in search of function. *Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun.* 66, 1148-1152. <https://doi.org/10.1107/S1744309110001685>
- Bellay, J., Han, S., Michaut, M., Kim, T., Costanzo, M., Andrews, B.J., Boone, C., Bader, G.D., Myers, C.L., Kim, P.M., 2011. Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol.* 12, R14. <https://doi.org/10.1186/gb-2011-12-2-r14>
- Benros, C., de Brevern, A.G., Etchebest, C., Hazout, S., 2006. Assessing a novel approach for predicting local 3D protein structures from sequence. *Proteins* 62, 865-880. <https://doi.org/10.1002/prot.20815>
- Bepler, T., Berger, B., 2019. Learning protein sequence embeddings using information from structure. <https://doi.org/10.48550/arXiv.1902.08661>
- Berman, H., Henrick, K., Nakamura, H., 2003. Announcing the worldwide Protein Data Bank. *Nat. Struct. Mol. Biol.* 10, 980-980. <https://doi.org/10.1038/nsb1203-980>
- Bernard, G., Pathmanathan, J.S., Lannes, R., Lopez, P., Baptiste, E., 2018. Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. *Genome Biol. Evol.* 10, 707-715. <https://doi.org/10.1093/gbe/evy031>
- Bershtein, S., Kleiner, D., Mishmar, D., 2021. Predicting 3D protein structures in light of evolution. *Nat. Ecol. Evol.* 5, 1195-1198. <https://doi.org/10.1038/s41559-021-01519-8>
- Bhowmick, A., Brookes, D.H., Yost, S.R., Dyson, H.J., Forman-Kay, J.D., Gunter, D., Head-Gordon, M., et al., 2016. Finding Our Way in the Dark Proteome. *J. Am. Chem. Soc.* 138, 9730-9742. <https://doi.org/10.1021/jacs.6b06543>
- Binder, J.L., Berendzen, J., Stevens, A.O., He, Y., Wang, J., Dokholyan, N.V., Oprea, T.I., 2021. AlphaFold Models Illuminate Half of Dark Human Proteins. <https://doi.org/10.1101/2021.11.04.467322>
- Bitard-Feildel, T., Callebaut, I., 2017. Exploring the dark foldable proteome by considering hydrophobic amino acids topology. *Sci. Rep.* 7, 41425. <https://doi.org/10.1038/srep41425>
- Bitard-Feildel, T., Heberlein, M., Bornberg-Bauer, E., Callebaut, I., 2015. Detection of Orphan Domains in *Drosophila* using «Hydrophobic Cluster Analysis». *Biochimie* 119. <https://doi.org/10.1016/j.biochi.2015.02.019>
- Bitard-Feildel, T., Lamiable, A., Mornon, J.-P., Callebaut, I., 2018. Order in Disorder as Observed by the «Hydrophobic Cluster Analysis» of Protein Sequences. *Proteomics* 18, e1800054. <https://doi.org/10.1002/pmic.201800054>
- Bittrich, S., Schroeder, M., Labudde, D., 2019. StructureDistiller: Structural relevance scoring identifies the most informative entries of a contact map. *Sci. Rep.* 9. <https://doi.org/10.1038/s41598-019-55047-4>
- Blake, J.D., Cohen, F.E., 2001. Pairwise sequence alignment below the twilight zone. Edited by B. Honig. *J. Mol. Biol.* 307, 721-735. <https://doi.org/10.1006/jmbi.2001.4495>
- Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasamy, S., Mitchell, A., Nuka, G., et al., 2021. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49, D344-D354. <https://doi.org/10.1093/nar/gkaa977>
- Bode, A.M., Dong, Z., 2004. Post-translational modification of p53 in tumorigenesis. *Nat. Rev. Cancer* 4, 793-805. <https://doi.org/10.1038/nrc1455>
- Bondarenko, V., Wells, M.M., Chen, Q., Tillman, T.S., Singewald, K., Lawless, M.J., Caporoso, J., et al., 2022. Structures of highly flexible intracellular domain of human  $\alpha 7$  nicotinic acetylcholine receptor. *Nat. Commun.* 13, 793. <https://doi.org/10.1038/s41467-022-28400-x>
- Bordin, N., Sillitoe, I., Lees, J.G., Orengo, C., 2021. Tracing Evolution Through Protein Structures: Nature Captured in a Few Thousand Folds. *Front. Mol. Biosci.* 8.

- Borgia, A., Borgia, M.B., Bugge, K., Kissling, V.M., Heidarsson, P.O., Fernandes, C.B., Sottini, A., et al., 2018. Extreme disorder in an ultrahigh-affinity protein complex. *Nature* 555, 61-66. <https://doi.org/10.1038/nature25762>
- Branden, C., Tooze, J., 1996. Introduction à la structure des protéines. De Boeck.
- Brown, C.J., Johnson, A.K., Dunker, A.K., Daughdrill, G.W., 2011. Evolution and Disorder. *Curr. Opin. Struct. Biol.* 21, 441-446. <https://doi.org/10.1016/j.sbi.2011.02.005>
- Brown, S.D., Babbitt, P.C., 2014. New Insights about Enzyme Evolution from Large Scale Studies of Sequence and Structure Relationships. *J. Biol. Chem.* 289, 30221-30228. <https://doi.org/10.1074/jbc.R114.569350>
- Buchan, D.W., Jones, D.T., 2017. EigenTHREADER: Analogous protein fold recognition by efficient contact map threading. *Bioinformatics* 33, 2684-2690.
- Buchan, D.W.A., Jones, D.T., 2019. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.* 47, W402-W407. <https://doi.org/10.1093/nar/gkz297>
- Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59-60. <https://doi.org/10.1038/nmeth.3176>
- Bugge, K., Brakti, I., Fernandes, C.B., Dreier, J.E., Lundsgaard, J.E., Olsen, J.G., Skriver, K., Kragelund, B.B., 2020. Interactions by Disorder – A Matter of Context. *Front. Mol. Biosci.* 7.
- Bungard, D., Copple, J.S., Yan, J., Chhun, J.J., Kumirov, V.K., Foy, S.G., Masel, J., Wysocki, V.H., Cordes, M.H.J., 2017. Foldability of a Natural De Novo Evolved Protein. *Structure* 25, 1687-1696.e4. <https://doi.org/10.1016/j.str.2017.09.006>
- Callaway, E., 2022. ‘The entire protein universe’: AI predicts shape of nearly every known protein. *Nature* 608, 15-16. <https://doi.org/10.1038/d41586-022-02083-2>
- Callebaut, I., Courvalin, J.C., Mornon, J.P., 1999. The BAH (bromo-adjacent homology) domain: a link between DNA methylation, replication and transcriptional regulation. *FEBS Lett.* 446, 189-193. [https://doi.org/10.1016/s0014-5793\(99\)00132-5](https://doi.org/10.1016/s0014-5793(99)00132-5)
- Callebaut, I., de Gunzburg, J., Goud, B., Mornon, J.P., 2001. RUN domains: a new family of domains involved in Ras-like GTPase signaling. *Trends Biochem. Sci.* 26, 79-83. [https://doi.org/10.1016/s0968-0004\(00\)01730-8](https://doi.org/10.1016/s0968-0004(00)01730-8)
- Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B., Mornon, J.P., 1997. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell. Mol. Life Sci. CMLS* 53, 621-645. <https://doi.org/10.1007/s000180050082>
- Callebaut, I., Mornon, J.-P., 2010. LOTUS, a new domain associated with small RNA pathways in the germline. *Bioinforma. Oxf. Engl.* 26, 1140-1144. <https://doi.org/10.1093/bioinformatics/btq122>
- Callebaut, I., Mornon, J.P., 1997. From BRCA1 to RAP1: a widespread BRCT module closely associated with DNA repair. *FEBS Lett.* 400, 25-30. [https://doi.org/10.1016/s0014-5793\(96\)01312-9](https://doi.org/10.1016/s0014-5793(96)01312-9)
- Callebaut, I., Moshous, D., Mornon, J.-P., de Villartay, J.-P., 2002. Metallo-beta-lactamase fold within nucleic acids processing enzymes: the beta-CASP family. *Nucleic Acids Res.* 30, 3592-3601. <https://doi.org/10.1093/nar/gkf470>
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421. <https://doi.org/10.1186/1471-2105-10-421>
- Camproux, A.C., Tuffery, P., Buffat, L., André, C., Boisvieux, J.F., Hazout, S., 1999. Analyzing patterns between regular secondary structures using short structural building blocks defined by a hidden Markov model. *Theor. Chem. Acc.* 101, 33-40. <https://doi.org/10.1007/s002140050402>
- Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., et al., 2018. A global ocean atlas of eukaryotic genes. *Nat. Commun.* 9, 373. <https://doi.org/10.1038/s41467-017-02342-1>
- Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charleatoux, B., et al., 2012. Proto-genes and de novo gene birth. *Nature* 487, 370-374. <https://doi.org/10.1038/nature11184>
- Chandonia, J.-M., Guan, L., Lin, S., Yu, C., Fox, N.K., Brenner, S.E., 2022. SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation



- and machine learning. *Nucleic Acids Res.* 50, D553-D559. <https://doi.org/10.1093/nar/gkab1054>
- Chemes, L.B., Alonso, L.G., Noval, M.G., de Prat-Gay, G., 2012. Circular dichroism techniques for the analysis of intrinsically disordered proteins and domains. *Methods Mol. Biol.* Clifton NJ 895, 387-404. [https://doi.org/10.1007/978-1-61779-927-3\\_22](https://doi.org/10.1007/978-1-61779-927-3_22)
- Chen, J.W., Romero, P., Uversky, V.N., Dunker, A.K., 2006. Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder. *J. Proteome Res.* 5, 888-898. <https://doi.org/10.1021/pr060049p>
- Cheng, H., Schaeffer, R.D., Liao, Y., Kinch, L.N., Pei, J., Shi, S., Kim, B.-H., Grishin, N.V., 2014. ECOD: An Evolutionary Classification of Protein Domains. *PLOS Comput. Biol.* 10, e1003926. <https://doi.org/10.1371/journal.pcbi.1003926>
- Chong, P.A., Forman-Kay, J.D., 2016. Liquid-liquid phase separation in cellular signaling systems. *Curr. Opin. Struct. Biol., Multi-protein assemblies in signaling • Catalysis and regulation* 41, 180-186. <https://doi.org/10.1016/j.sbi.2016.08.001>
- Chong, S., Mir, M., 2021. Towards Decoding the Sequence-Based Grammar Governing the Functions of Intrinsically Disordered Protein Regions. *J. Mol. Biol., Phase Separation in Biology and Disease: The Next Chapter* 433, 166724. <https://doi.org/10.1016/j.jmb.2020.11.023>
- Chothia, C., 1992. One thousand families for the molecular biologist. *Nature* 357, 543-544. <https://doi.org/10.1038/357543a0>
- Chowdhury, R., Bouatta, N., Biswas, S., Rochereau, C., Church, G.M., Sorger, P.K., AlQuraishi, M., 2021. Single-sequence protein structure prediction using language models from deep learning. <https://doi.org/10.1101/2021.08.02.454840>
- Cioce, M., Lamond, A.I., 2005. Cajal bodies: a long history of discovery. *Annu. Rev. Cell Dev. Biol.* 21, 105-131. <https://doi.org/10.1146/annurev.cellbio.20.010403.103738>
- Cretin, G., Galochkina, T., Vander Meersche, Y., de Brevern, A.G., Postic, G., Gelly, J.-C., 2022. SWORD2: hierarchical analysis of protein 3D structures. *Nucleic Acids Res.* 50, W732-W738. <https://doi.org/10.1093/nar/gkac370>
- Daniels, A.J., Williams, R.J.P., Wright, P.E., 1978. The character of the stored molecules in chromaffin granules of the adrenal medulla: A nuclear magnetic resonance study. *Neuroscience* 3, 573-585. [https://doi.org/10.1016/0306-4522\(78\)90022-2](https://doi.org/10.1016/0306-4522(78)90022-2)
- Dass, R., Mulder, F.A.A., Nielsen, J.T., 2020. ODINPred: comprehensive prediction of protein order and disorder. *Sci. Rep.* 10, 14780. <https://doi.org/10.1038/s41598-020-71716-1>
- Davey, N.E., Roey, K.V., Weatheritt, R.J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., Gibson, T.J., 2011. Attributes of short linear motifs. *Mol. Biosyst.* 8, 268-281. <https://doi.org/10.1039/C1MB05231D>
- de Brevern, A.G., Etchebest, C., Hazout, S., 2000. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins* 41, 271-287. [https://doi.org/10.1002/1097-0134\(20001115\)41:3<271::aid-prot10>3.0.co;2-z](https://doi.org/10.1002/1097-0134(20001115)41:3<271::aid-prot10>3.0.co;2-z)
- Diella, F., Haslam, N., Chica, C., Budd, A., Michael, S., Brown, N.P., Trave, G., Gibson, T.J., 2008. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front. Biosci. J. Virtual Libr.* 13, 6580-6603. <https://doi.org/10.2741/3175>
- Dimonaco, N.J., Aubrey, W., Kenobi, K., Clare, A., Creevey, C.J., 2022. No one tool to rule them all: prokaryotic gene prediction tool annotations are highly dependent on the organism of study. *Bioinformatics* 38, 1198-1207. <https://doi.org/10.1093/bioinformatics/btab827>
- Dinkel, H., Van Roey, K., Michael, S., Davey, N.E., Weatheritt, R.J., Born, D., Speck, T., et al., 2014. The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Res.* 42, D259-266. <https://doi.org/10.1093/nar/gkt1047>
- Dong, Q., Wang, X., Lin, L., 2008. Prediction of protein local structures and folding fragments based on building-block library. *Proteins* 72, 353-366. <https://doi.org/10.1002/prot.21931>
- Dosztányi, Z., Csizmók, V., Tompa, P., Simon, I., 2005a. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* 347, 827-839. <https://doi.org/10.1016/j.jmb.2005.01.071>
- Dosztányi, Z., Csizmók, V., Tompa, P., Simon, I., 2005b. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21, 3433-3434. <https://doi.org/10.1093/bioinformatics/bti541>

- Dreier, J.E., Prestel, A., Martins, J.M., Brøndum, S.S., Nielsen, O., Garbers, A.E., Suga, H., et al., 2022. A context-dependent and disordered ubiquitin-binding motif. *Cell. Mol. Life Sci.* 79, 484. <https://doi.org/10.1007/s00018-022-04486-w>
- Drozdetskiy, A., Cole, C., Procter, J., Barton, G.J., 2015. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* 43, W389-W394. <https://doi.org/10.1093/nar/gkv332>
- Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., et al., 2001. Intrinsically disordered protein. *J. Mol. Graph. Model.* 19, 26-59. [https://doi.org/10.1016/s1093-3263\(00\)00138-8](https://doi.org/10.1016/s1093-3263(00)00138-8)
- Duval, M., Cossart, P., 2017. Small bacterial and phagic proteins: an updated view on a rapidly moving field. *Curr. Opin. Microbiol., Antimicrobials \* Bacterial Systems Biology* 39, 81-88. <https://doi.org/10.1016/j.mib.2017.09.010>
- Dyson, H.J., Wright, P.E., 2004. Unfolded proteins and protein folding studied by NMR. *Chem. Rev.* 104, 3607-3622. <https://doi.org/10.1021/cr030403s>
- Dyson, H.J., Wright, P.E., 1998. Equilibrium NMR studies of unfolded and partially folded proteins. *Nat. Struct. Biol.* 5, 499-503. <https://doi.org/10.1038/739>
- Ebbert, M.T.W., Jensen, T.D., Jansen-West, K., Sens, J.P., Reddy, J.S., Ridge, P.G., Kauwe, J.S.K., et al., 2019. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol.* 20, 97. <https://doi.org/10.1186/s13059-019-1707-2>
- Eberhardt, R.Y., Haft, D.H., Punta, M., Martin, M., O'Donovan, C., Bateman, A., 2012. AntiFam: a tool to help identify spurious ORFs in protein annotation. *Database J. Biol. Databases Curation* 2012, bas003. <https://doi.org/10.1093/database/bas003>
- Eddy, S.R., 2011. Accelerated Profile HMM Searches. *PLOS Comput. Biol.* 7, e1002195. <https://doi.org/10.1371/journal.pcbi.1002195>
- Ekman, D., Elofsson, A., 2010. Identifying and quantifying orphan protein sequences in fungi. *J. Mol. Biol.* 396, 396-405. <https://doi.org/10.1016/j.jmb.2009.11.053>
- Ellens, K.W., Christian, N., Singh, C., Satagopam, V.P., May, P., Linster, C.L., 2017. Confronting the catalytic dark matter encoded by sequenced genomes. *Nucleic Acids Res.* 45, 11495-11514. <https://doi.org/10.1093/nar/gkx937>
- Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., et al., 2022. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 7112-7127. <https://doi.org/10.1109/tpami.2021.3095381>
- Eudes, R., Le Tuan, K., Delettré, J., Mornon, J.-P., Callebaut, I., 2007. A generalized analysis of hydrophobic and loop clusters within globular protein sequences. *BMC Struct. Biol.* 7, 2. <https://doi.org/10.1186/1472-6807-7-2>
- Faure, G., Callebaut, I., 2013. Comprehensive repertoire of foldable regions within whole genomes. *PLoS Comput. Biol.* 9, e1003280. <https://doi.org/10.1371/journal.pcbi.1003280>
- Fichó, E., Reményi, I., Simon, I., Mészáros, B., 2017. MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinforma. Oxf. Engl.* 33. <https://doi.org/10.1093/bioinformatics/btx486>
- Fischer, S., Brunk, B.P., Chen, F., Gao, X., Harb, O.S., Iodice, J.B., Shanmugam, D., Roos, D.S., Stoeckert, C.J., 2011. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr. Protoc. Bioinforma.* Chapter 6, Unit 6.12.1-19. <https://doi.org/10.1002/0471250953.bi0612s35>
- Flock, T., Weatheritt, R.J., Latysheva, N.S., Babu, M.M., 2014. Controlling entropy to tune the functions of intrinsically disordered regions. *Curr. Opin. Struct. Biol.* 26, 62-72. <https://doi.org/10.1016/j.sbi.2014.05.007>
- Fourty, G., Callebaut, I., Mornon, J.-P., 2008. Characterization of non-trivial neighborhood fold constraints from protein sequences using generalized topohydrophobicity. *Bioinforma. Biol. Insights* 2, 47-66. <https://doi.org/10.4137/bbi.s426>
- Fowler, D.M., Fields, S., 2014. Deep mutational scanning: a new style of protein science. *Nat. Methods* 11, 801-807. <https://doi.org/10.1038/nmeth.3027>
- Fox, N.K., Brenner, S.E., Chandonia, J.-M., 2014. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* 42, D304-D309. <https://doi.org/10.1093/nar/gkt1240>

- Fuxreiter, M., Simon, I., Friedrich, P., Tompa, P., 2004. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.* 338, 1015-1026. <https://doi.org/10.1016/j.jmb.2004.03.017>
- Fuxreiter, M., Tompa, P., Simon, I., 2007. Local structural disorder imparts plasticity on linear motifs. *Bioinforma. Oxf. Engl.* 23, 950-956. <https://doi.org/10.1093/bioinformatics/btm035>
- Gabaldón, T., Koonin, E.V., 2013. Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* 14, 360-366. <https://doi.org/10.1038/nrg3456>
- Gaboriaud, C., Bissery, V., Benchetrit, T., Mornon, J.P., 1987. Hydrophobic cluster analysis: An efficient new way to compare and analyse amino acid sequences. *FEBS Lett.* 224, 149-155. [https://doi.org/10.1016/0014-5793\(87\)80439-8](https://doi.org/10.1016/0014-5793(87)80439-8)
- Galea, C.A., Wang, Y., Sivakolundu, S.G., Kriwacki, R.W., 2008. Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. *Biochemistry* 47, 7598-7609. <https://doi.org/10.1021/bi8006803>
- Galperin, M.Y., Koonin, E.V., 2004. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res.* 32, 5452-5463. <https://doi.org/10.1093/nar/gkh885>
- Galzitskaya, O.V., Garbuzynskiy, S.O., Lobanov, M.Yu., 2006. FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics* 22, 2948-2949. <https://doi.org/10.1093/bioinformatics/btl504>
- Ghouzam, Y., Postic, G., de Brevern, A.G., Gelly, J.-C., 2015. Improving protein fold recognition with hybrid profiles combining sequence and structure evolution. *Bioinforma. Oxf. Engl.* 31, 3782-3789. <https://doi.org/10.1093/bioinformatics/btv462>
- Godzik, A., 2011. Metagenomics and the protein universe. *Curr. Opin. Struct. Biol.* 21, 398-403. <https://doi.org/10.1016/j.sbi.2011.03.010>
- Goodacre, N.F., Gerloff, D.L., Uetz, P., 2013. Protein domains of unknown function are essential in bacteria. *mBio* 5, e00744-00713. <https://doi.org/10.1128/mBio.00744-13>
- Goodwin, S., McPherson, J.D., McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333-351. <https://doi.org/10.1038/nrg.2016.49>
- Greener, J.G., Kandathil, S.M., Jones, D.T., 2019. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat. Commun.* 10. <https://doi.org/10.1038/s41467-019-11994-0>
- Gsponer, J., Babu, M.M., 2009. The rules of disorder or why disorder rules. *Prog. Biophys. Mol. Biol.* 99, 94-103. <https://doi.org/10.1016/j.pbiomolbio.2009.03.001>
- Habchi, J., Tompa, P., Longhi, S., Uversky, V.N., 2014. Introducing Protein Intrinsic Disorder. *Chem. Rev.* 114, 6561-6588. <https://doi.org/10.1021/cr400514h>
- Hamilton, W.L., Claessens, A., Otto, T.D., Kekre, M., Fairhurst, R.M., Rayner, J.C., Kwiatkowski, D., 2017. Extreme mutation bias and high AT content in *Plasmodium falciparum*. *Nucleic Acids Res.* 45, 1889-1901. <https://doi.org/10.1093/nar/gkw1259>
- Hanson, J., Paliwal, K., Zhou, Y., 2018. Accurate Single-Sequence Prediction of Protein Intrinsic Disorder by an Ensemble of Deep Recurrent and Convolutional Architectures. *J. Chem. Inf. Model.* 58, 2369-2376. <https://doi.org/10.1021/acs.jcim.8b00636>
- Hanson, J., Paliwal, K.K., Litfin, T., Zhou, Y., 2019. SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning. *Genomics Proteomics Bioinformatics* 17, 645-656. <https://doi.org/10.1016/j.gpb.2019.01.004>
- Hatos, A., Hajdu-Soltész, B., Monzon, A.M., Palopoli, N., Álvarez, L., Aykac-Fas, B., Bassot, C., et al., 2020. DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* 48, D269-D276. <https://doi.org/10.1093/nar/gkz975>
- Heinzinger, M., Littmann, M., Sillitoe, I., Bordin, N., Orengo, C., Rost, B., 2022. Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genomics Bioinforma.* 4, lqac043. <https://doi.org/10.1093/nargab/lqac043>
- Hennetin, J., Le, T., Canard, L., Colloc'h, N., Mornon, J.-P., Callebaut, I., 2003. Non-intertwined binary patterns of hydrophobic/nonhydrophobic amino acids are considerably better markers of regular secondary structures than nonconstrained patterns. *Proteins* 51, 236-44. <https://doi.org/10.1002/prot.10355>
- Holm, L., 2020. DALI and the persistence of protein shape. *Protein Sci.* 29, 128-140. <https://doi.org/10.1002/pro.3749>

- Holm, L., Sander, C., 1996. Mapping the protein universe. *Science* 273, 595-603. <https://doi.org/10.1126/science.273.5275.595>
- Hong, H., 2014. Toward understanding driving forces in membrane protein folding. *Arch. Biochem. Biophys.* 564, 297-313. <https://doi.org/10.1016/j.abb.2014.07.031>
- Höps, W., Jeffryes, M., Bateman, A., 2018. Gene Unprediction with Spurio: A tool to identify spurious protein sequences. *F1000Research* 7, 261. <https://doi.org/10.12688/f1000research.14050.1>
- Hu, G., Katuwawala, A., Wang, K., Wu, Z., Ghadermarzi, S., Gao, J., Kurgan, L., 2021. fDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nat. Commun.* 12, 4438. <https://doi.org/10.1038/s41467-021-24773-7>
- Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., Bork, P., 2017. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* 34, 2115-2122. <https://doi.org/10.1093/molbev/msx148>
- Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., et al., 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309-D314. <https://doi.org/10.1093/nar/gky1085>
- Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., et al., 2016. A new view of the tree of life. *Nat. Microbiol.* 1, 16048. <https://doi.org/10.1038/nmicrobiol.2016.48>
- Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11, 119. <https://doi.org/10.1186/1471-2105-11-119>
- Jakob, U., Kriwacki, R., Uversky, V.N., 2014. Conditionally and transiently disordered proteins: awakening cryptic disorder to regulate protein function. *Chem. Rev.* 114, 6779-6805. <https://doi.org/10.1021/cr400459c>
- Jaroszewski, L., Li, Z., Krishna, S.S., Bakolitsa, C., Wooley, J., Deacon, A.M., Wilson, I.A., Godzik, A., 2009. Exploration of Uncharted Regions of the Protein Universe. *PLOS Biol.* 7, e1000205. <https://doi.org/10.1371/journal.pbio.1000205>
- Jasinska, W., Manhart, M., Lerner, J., Gauthier, L., Serohijos, A.W.R., Bershtein, S., 2020. Chromosomal barcoding of *E. coli* populations reveals lineage diversity dynamics at high resolution. *Nat. Ecol. Evol.* 4, 437-452. <https://doi.org/10.1038/s41559-020-1103-z>
- Jiang, Y., Oron, T.R., Clark, W.T., Bankapur, A.R., D'Andrea, D., Lepore, R., Funk, C.S., et al., 2016. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.* 17, 184. <https://doi.org/10.1186/s13059-016-1037-6>
- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., Charpentier, E., 2012. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* 337, 816-821. <https://doi.org/10.1126/science.1225829>
- Jisna, V.A., Jayaraj, P.B., 2021. Protein Structure Prediction: Conventional and Deep Learning Perspectives. *Protein J.* 40, 522-544. <https://doi.org/10.1007/s10930-021-10003-y>
- Jones, C.L., Tepe, J.J., 2019. Proteasome Activation to Combat Proteotoxicity. *Molecules* 24, 2841. <https://doi.org/10.3390/molecules24152841>
- Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., et al., 2014. InterProScan 5: genome-scale protein function classification. *Bioinforma. Oxf. Engl.* 30, 1236-1240. <https://doi.org/10.1093/bioinformatics/btu031>
- Joseph, A.P., Agarwal, G., Mahajan, S., Gelly, J.-C., Swapna, L.S., Offmann, B., Cadet, F., et al., 2010. A short survey on protein blocks. *Biophys. Rev.* 2, 137-145. <https://doi.org/10.1007/s12551-010-0036-1>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., et al., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583-589. <https://doi.org/10.1038/s41586-021-03819-2>
- Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577-2637. <https://doi.org/10.1002/bip.360221211>
- Kandathil, S.M., Greener, J.G., Jones, D.T., 2019. Recent developments in deep learning applied to protein structure prediction. *Proteins* 87, 1179-1189. <https://doi.org/10.1002/prot.25824>

- Kato, M., Han, T.W., Xie, S., Shi, K., Du, X., Wu, L.C., Mirzaei, H., et al., 2012. Cell-free Formation of RNA Granules: Low Complexity Sequence Domains Form Dynamic Fibers within Hydrogels. *Cell* 149, 753-767. <https://doi.org/10.1016/j.cell.2012.04.017>
- Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., Sternberg, M.J.E., 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10, 845-858. <https://doi.org/10.1038/nprot.2015.053>
- Kempen, M. van, Kim, S.S., Tumescheit, C., Mirdita, M., Gilchrist, C.L.M., Söding, J., Steinegger, M., 2022. Foldseek: fast and accurate protein structure search. <https://doi.org/10.1101/2022.02.07.479398>
- Kodera, N., Ando, T., 2022. Guide to studying intrinsically disordered proteins by high-speed atomic force microscopy. *Methods* 207, 44-56. <https://doi.org/10.1016/j.ymeth.2022.08.008>
- Kolodny, R., Koehl, P., Guibas, L., Levitt, M., 2002. Small Libraries of Protein Fragments Model Native Protein Structures Accurately. *J. Mol. Biol.* 323, 297-307. [https://doi.org/10.1016/S0022-2836\(02\)00942-7](https://doi.org/10.1016/S0022-2836(02)00942-7)
- Kolodny, R., Pereyaslavets, L., Samson, A.O., Levitt, M., 2013. On the Universe of Protein Folds. *Annu. Rev. Biophys.* 42, 559-582. <https://doi.org/10.1146/annurev-biophys-083012-130432>
- Koonin, E.V., 2005. Orthologs, Paralogs, and Evolutionary Genomics. *Annu. Rev. Genet.* 39, 309-338. <https://doi.org/10.1146/annurev.genet.39.073003.114725>
- Koonin, E.V., Wolf, Y.I., Karev, G.P., 2002. The structure of the protein universe and genome evolution. *Nature* 420, 218-223. <https://doi.org/10.1038/nature01256>
- Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K., Fernandez-Guerra, A., et al., 2015. The ocean sampling day consortium. *GigaScience* 4, 27. <https://doi.org/10.1186/s13742-015-0066-5>
- Kouzarides, T., 2007. Chromatin modifications and their function. *Cell* 128, 693-705. <https://doi.org/10.1016/j.cell.2007.02.005>
- Ku, S.-Y., Hu, Y.-J., 2008. Protein structure search and local structure characterization. *BMC Bioinformatics* 9, 349. <https://doi.org/10.1186/1471-2105-9-349>
- La, D., Esquivel-Rodríguez, J., Venkatraman, V., Li, B., Sael, L., Ueng, S., Ahrendt, S., Kihara, D., 2009. 3D-SURFER: software for high-throughput protein surface comparison and analysis. *Bioinforma. Oxf. Engl.* 25, 2843-2844. <https://doi.org/10.1093/bioinformatics/btp542>
- Labesse, G., Colloc'h, N., Pothier, J., Mornon, J.-P., 1997. P-SEA: a new efficient assignment of secondary structure from C $\alpha$  trace of proteins. *Bioinformatics* 13, 291-295. <https://doi.org/10.1093/bioinformatics/13.3.291>
- Laine, E., Eismann, S., Eloffsson, A., Grudin, S., 2021. Protein sequence-to-structure learning: Is this the end(-to-end revolution)?
- Lamiable, A., Bitard-Feildel, T., Rebehmed, J., Quintus, F., Schoentgen, F., Mornon, J.-P., Callebaut, I., 2019. A topology-based investigation of protein interaction sites using Hydrophobic Cluster Analysis. *Biochimie* 167, 68-80. <https://doi.org/10.1016/j.biochi.2019.09.009>
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., et al., 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860-921. <https://doi.org/10.1038/35057062>
- Leduc, V., 2006. Étude de l'interchangeabilité des amas hydrophobes locaux au sein des domaines globulaires protéiques.
- Levin, L., Mishmar, D., 2017. The genomic landscape of evolutionary convergence in mammals, birds and reptiles. *Nat. Ecol. Evol.* 1, 0041. <https://doi.org/10.1038/s41559-016-0041>
- Li, W., O'Neill, K.R., Haft, D.H., DiCuccio, M., Chetvernin, V., Badretdin, A., Coulouris, G., et al., 2021. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res.* 49, D1020-D1028. <https://doi.org/10.1093/nar/gkaa1105>
- Liljas, A., Liljas, L., Nissen, P., Lindblom, G., Ash, M.-R., 2017. *Textbook of Structural Biology*. World Scientific.
- Linding, R., Russell, R.B., Neduva, V., Gibson, T.J., 2003. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* 31, 3701-3708. <https://doi.org/10.1093/nar/gkg519>

- Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T., Rost, B., 2021. Embeddings from deep learning transfer GO annotations beyond homology. *Sci. Rep.* 11, 1160. <https://doi.org/10.1038/s41598-020-80786-0>
- Liu, Y., Wang, X., Liu, B., 2019. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief. Bioinform.* 20, 330-346. <https://doi.org/10.1093/bib/bbx126>
- Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A.B., Brady, A., et al., 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* 550, 61-66. <https://doi.org/10.1038/nature23889>
- Lomize, M.A., Lomize, A.L., Pogozheva, I.D., Mosberg, H.I., 2006. OPM: Orientations of Proteins in Membranes database. *Bioinformatics* 22, 623-625. <https://doi.org/10.1093/bioinformatics/btk023>
- Lu, T.-C., Leu, J.-Y., Lin, W.-C., 2017. A Comprehensive Analysis of Transcript-Supported De Novo Genes in *Saccharomyces sensu stricto* Yeasts. *Mol. Biol. Evol.* 34, 2823-2838. <https://doi.org/10.1093/molbev/msx210>
- Mahlich, Y., Steinegger, M., Rost, B., Bromberg, Y., 2018. HFSP: high speed homology-driven function annotation of proteins. *Bioinforma. Oxf. Engl.* 34, i304-i312. <https://doi.org/10.1093/bioinformatics/bty262>
- Marchler-Bauer, A., Bryant, S.H., 2004. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* 32, W327-331. <https://doi.org/10.1093/nar/gkh454>
- Mariani, V., Biasini, M., Barbato, A., Schwede, T., 2013. IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* 29, 2722-2728. <https://doi.org/10.1093/bioinformatics/btt473>
- Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., Sander, C., 2011. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLOS ONE* 6, e28766. <https://doi.org/10.1371/journal.pone.0028766>
- Méheust, R., Castelle, C.J., Jaffe, A.L., Banfield, J.F., 2022. Conserved and lineage-specific hypothetical proteins may have played a central role in the rise and diversification of major archaeal groups. *BMC Biol.* 20, 154. <https://doi.org/10.1186/s12915-022-01348-6>
- Melnyk, R.A., Hossain, S.S., Haney, C.H., 2019. Convergent gain and loss of genomic islands drive lifestyle changes in plant-associated *Pseudomonas*. *ISME J.* 13, 1575-1588. <https://doi.org/10.1038/s41396-019-0372-5>
- Mészáros, B., Erdos, G., Dosztányi, Z., 2018. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* 46, W329-W337. <https://doi.org/10.1093/nar/gky384>
- Mi, T., Merlin, J.C., Deverasetty, S., Gryk, M.R., Bill, T.J., Brooks, A.W., Lee, L.Y., et al., 2012. MinMotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. *Nucleic Acids Res.* 40, D252-D260. <https://doi.org/10.1093/nar/gkr1189>
- Mier, P., Paladin, L., Tamana, S., Petrosian, S., Hajdu-Soltész, B., Urbanek, A., Gruca, A., et al., 2020. Disentangling the complexity of low complexity proteins. *Brief. Bioinform.* 21, 458-472. <https://doi.org/10.1093/bib/bbz007>
- Miller, D., Stern, A., Burstein, D., 2022. Deciphering microbial gene function using natural language processing. *Nat. Commun.* 13, 5731. <https://doi.org/10.1038/s41467-022-33397-4>
- Mirabello, C., Wallner, B., 2019. rawMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments. *PLOS ONE* 14, e0220182. <https://doi.org/10.1371/journal.pone.0220182>
- Misikei, M., Horvath, A., Vendruscolo, M., Fuxreiter, M., 2020. Sequence-Based Prediction of Fuzzy Protein Interactions. *J. Mol. Biol.* 432, 2289-2303. <https://doi.org/10.1016/j.jmb.2020.02.017>
- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., et al., 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49, D412-D419. <https://doi.org/10.1093/nar/gkaa913>
- Mizuguchi, K., Go, N., 1995. Comparison of spatial arrangements of secondary structural elements in proteins. *Protein Eng.* 8, 353-362. <https://doi.org/10.1093/protein/8.4.353>

- Mohan, A., Oldfield, C.J., Radivojac, P., Vacic, V., Cortese, M.S., Dunker, A.K., Uversky, V.N., 2006. Analysis of Molecular Recognition Features (MoRFs). *J. Mol. Biol.* 362, 1043-1059. <https://doi.org/10.1016/j.jmb.2006.07.087>
- Monzon, V., Haft, D.H., Bateman, A., 2022a. Folding the unfoldable: using AlphaFold to explore spurious proteins. *Bioinforma. Adv.* 2, vbab043. <https://doi.org/10.1093/bioadv/vbab043>
- Monzon, V., Paysan-Lafosse, T., Wood, V., Bateman, A., 2022b. Reciprocal Best Structure Hits: Using AlphaFold models to discover distant homologues. <https://doi.org/10.1101/2022.07.04.498216>
- Mudgal, R., Sandhya, S., Chandra, N., Srinivasan, N., 2015. De-DUFing the DUFs: Deciphering distant evolutionary relationships of Domains of Unknown Function using sensitive homology detection methods. *Biol. Direct* 10, 38. <https://doi.org/10.1186/s13062-015-0069-2>
- Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., et al., 2021. A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* 39, 499-509. <https://doi.org/10.1038/s41587-020-0718-6>
- Necci, M., Piovesan, D., Tosatto, S.C.E., 2021. Critical assessment of protein intrinsic disorder prediction. *Nat. Methods* 18, 472-481. <https://doi.org/10.1038/s41592-021-01117-3>
- Ng, J.C.-F., Fraternali, F., 2021. Protein structural consequences of DNA mutational signatures: A meta-analysis of somatic variants and deep mutational scanning data. <https://doi.org/10.1101/2021.05.27.445950>
- O'Donoghue, S.I., Sabir, K.S., Kalemanov, M., Stolte, C., Wellmann, B., Ho, V., Roos, M., et al., 2015. Aquaria: simplifying discovery and insight from protein structures. *Nat. Methods* 12, 98-99. <https://doi.org/10.1038/nmeth.3258>
- Oldfield, C.J., Cheng, Y., Cortese, M.S., Romero, P., Uversky, V.N., Dunker, A.K., 2005. Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* 44, 12454-12470. <https://doi.org/10.1021/bi050736e>
- Oss, S.B.V., Carvunis, A.-R., 2019. De novo gene birth. *PLOS Genet.* 15, e1008160. <https://doi.org/10.1371/journal.pgen.1008160>
- Outeiral, C., Nissley, D.A., Deane, C.M., 2022. Current structure predictors are not learning the physics of protein folding. *Bioinformatics* 38, 1881-1887. <https://doi.org/10.1093/bioinformatics/btab881>
- Pace, C.N., Scholtz, J.M., Grimsley, G.R., 2014. Forces stabilizing proteins. *FEBS Lett.* 588, 2177-2184. <https://doi.org/10.1016/j.febslet.2014.05.006>
- Papadopoulos, C., Callebaut, I., Gelly, J.-C., Hatin, I., Namy, O., Renard, M., Lespinet, O., Lopes, A., 2021. Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution. *Genome Res.* 31, 2303-2315. <https://doi.org/10.1101/gr.275638.121>
- Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., Hugenholtz, P., 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* 50, D785-D794. <https://doi.org/10.1093/nar/gkab776>
- Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., Hugenholtz, P., 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36, 996-1004. <https://doi.org/10.1038/nbt.4229>
- Pauling, L., Corey, R.B., 1951. Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets. *Proc. Natl. Acad. Sci. U. S. A.* 37, 729-740. <https://doi.org/10.1073/pnas.37.11.729>
- Peng, Z., Yan, J., Fan, X., Mizianty, M.J., Xue, B., Wang, K., Hu, G., Uversky, V.N., Kurgan, L., 2015. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell. Mol. Life Sci. CMLS* 72, 137-151. <https://doi.org/10.1007/s00018-014-1661-9>
- Pentony, M.M., Jones, D.T., 2010. Modularity of intrinsic disorder in the human proteome. *Proteins* 78, 212-221. <https://doi.org/10.1002/prot.22504>
- Perdigão, N., Heinrich, J., Stolte, C., Sabir, K.S., Buckley, M.J., Tabor, B., Signal, B., et al., 2015. Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci.* 112, 15898-15903. <https://doi.org/10.1073/pnas.1508380112>

- Pereira, J., Simpkin, A.J., Hartmann, M.D., Rigden, D.J., Keegan, R.M., Lupas, A.N., 2021. High-accuracy protein structure prediction in CASP14. *Proteins* 89, 1687-1699. <https://doi.org/10.1002/prot.26171>
- Piovesan, D., Monzon, A.M., Quaglia, F., Tosatto, S.C.E., 2022. Databases for intrinsically disordered proteins. *Acta Crystallogr. Sect. Struct. Biol.* 78, 144-151. <https://doi.org/10.1107/S2059798321012109>
- Piovesan, D., Necci, M., Escobedo, N., Monzon, A.M., Hatos, A., Mičetić, I., Quaglia, F., et al., 2020. MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.* 49, D361-D367. <https://doi.org/10.1093/nar/gkaa1058>
- Pizzi, E., Frontali, C., 2001. Low-complexity regions in *Plasmodium falciparum* proteins. *Genome Res.* 11, 218-229. <https://doi.org/10.1101/gr.gr-1522r>
- Porta-Pardo, E., Ruiz-Serra, V., Valentini, S., Valencia, A., 2022. The structural coverage of the human proteome before and after AlphaFold. *PLOS Comput. Biol.* 18, e1009818. <https://doi.org/10.1371/journal.pcbi.1009818>
- Poupon, A., Mornon, J.P., 1998. Populations of hydrophobic amino acids within protein globular domains: identification of conserved « topohydrophobic » positions. *Proteins* 33, 329-342. [https://doi.org/10.1002/\(sici\)1097-0134\(19981115\)33:3<329::aid-prot3>3.0.co;2-e](https://doi.org/10.1002/(sici)1097-0134(19981115)33:3<329::aid-prot3>3.0.co;2-e)
- Promponas, V.J., Enright, A.J., Tsoka, S., Kreil, D.P., Leroy, C., Hamodrakas, S., Sander, C., Ouzounis, C.A., 2000. CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. *Bioinforma. Oxf. Engl.* 16, 915-922. <https://doi.org/10.1093/bioinformatics/16.10.915>
- Rebehmed, J., Quintus, F., Mornon, J.-P., Callebaut, I., 2016a. The respective roles of polar/nonpolar binary patterns and amino acid composition in protein regular secondary structures explored exhaustively using hydrophobic cluster analysis. *Proteins Struct. Funct. Bioinforma.* 84, 624-638. <https://doi.org/10.1002/prot.25012>
- Rebehmed, J., Quintus, F., Mornon, J.-P., Callebaut, I., 2016b. 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, 624-638. <https://doi.org/10.1002/prot.25012>
- Receveur-Bréchet, V., Bourhis, J.-M., Uversky, V.N., Canard, B., Longhi, S., 2006. Assessing protein disorder and induced folding. *Proteins* 62, 24-45. <https://doi.org/10.1002/prot.20750>
- Redfern, O.C., Harrison, A., Dallman, T., Pearl, F.M.G., Orengo, C.A., 2007. CATHEDRAL: A Fast and Effective Algorithm to Predict Folds and Domain Boundaries from Multidomain Protein Structures. *PLOS Comput. Biol.* 3, e232. <https://doi.org/10.1371/journal.pcbi.0030232>
- Reed, D.C., Algar, C.K., Huber, J.A., Dick, G.J., 2014. Gene-centric approach to integrating environmental genomics and biogeochemical models. *Proc. Natl. Acad. Sci.* 111, 1879-1884. <https://doi.org/10.1073/pnas.1313713111>
- Remm, M., Storm, C.E., Sonnhammer, E.L., 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041-1052. <https://doi.org/10.1006/jmbi.2000.5197>
- Richards, F.M., Kundrot, C.E., 1988. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins* 3, 71-84. <https://doi.org/10.1002/prot.340030202>
- Riesenfeld, C.S., Goodman, R.M., Handelsman, J., 2004. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environ. Microbiol.* 6, 981-989. <https://doi.org/10.1111/j.1462-2920.2004.00664.x>
- Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., et al., 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431-437. <https://doi.org/10.1038/nature12352>
- Río, Á.R. del, Giner-Lamia, J., Cantalapiedra, C.P., Botas, J., Deng, Z., Hernández-Plaza, A., Paoli, L., et al., 2022. Functional and evolutionary significance of unknown genes from uncultivated taxa. <https://doi.org/10.1101/2022.01.26.477801>
- Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., et al., 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.* 118, e2016239118. <https://doi.org/10.1073/pnas.2016239118>



- Robbe-Saule, V., Lopes, M.D., Kolb, A., Norel, F., 2007. Physiological Effects of Crl in Salmonella Are Modulated by  $\sigma$ S Level and Promoter Specificity. *J. Bacteriol.* 189, 2976-2987. <https://doi.org/10.1128/JB.01919-06>
- Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J.E., Dunker, A.K., 1997. Identifying disordered regions in proteins from amino acid sequences, in: *IEEE Int. Conf. Neural Netw.* p. 90-95.
- Ruff, K.M., Pappu, R.V., 2021. AlphaFold and Implications for Intrinsically Disordered Proteins. *J. Mol. Biol., From Protein Sequence to Structure at Warp Speed: How AlphaFold Impacts Biology* 433, 167208. <https://doi.org/10.1016/j.jmb.2021.167208>
- Sahoo, S., Mahapatra, S.R., Das, N., Parida, B.K., Rath, S., Misra, N., Suar, M., 2020. Functional elucidation of hypothetical proteins associated with lipid accumulation: Prioritizing genetic engineering targets for improved algal biofuel production. *Algal Res.* 47, 101887. <https://doi.org/10.1016/j.algal.2020.101887>
- Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M., Smith, M., 1977. Nucleotide sequence of bacteriophage  $\phi$ X174 DNA. *Nature* 265, 687-695. <https://doi.org/10.1038/265687a0>
- Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., et al., 2021. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 50, D20-D26. <https://doi.org/10.1093/nar/gkab1112>
- Scaiewicz, A., Levitt, M., 2018. Unique function words characterize genomic proteins. *Proc. Natl. Acad. Sci.* 115, 6703-6708. <https://doi.org/10.1073/pnas.1801182115>
- Scaiewicz, A., Levitt, M., 2015. The language of the protein universe. *Curr. Opin. Genet. Dev.* 35, 50-56. <https://doi.org/10.1016/j.gde.2015.08.010>
- Schad, E., Fichó, E., Pancsa, R., Simon, I., Dosztányi, Z., Mészáros, B., 2018. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics* 34, 535-537. <https://doi.org/10.1093/bioinformatics/btx640>
- Schlessinger, A., Punta, M., Rost, B., 2007. Natively unstructured regions in proteins identified from contact predictions. *Bioinforma. Oxf. Engl.* 23, 2376-2384. <https://doi.org/10.1093/bioinformatics/btm349>
- Schlötterer, C., 2015. Genes from scratch--the evolutionary fate of de novo genes. *Trends Genet. TIG* 31, 215-219. <https://doi.org/10.1016/j.tig.2015.02.007>
- Schmitz, J.F., Ullrich, K.K., Bornberg-Bauer, E., 2018. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat. Ecol. Evol.* 2, 1626-1632. <https://doi.org/10.1038/s41559-018-0639-7>
- Schnoes, A.M., Brown, S.D., Dodevski, I., Babbitt, P.C., 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* 5, e1000605. <https://doi.org/10.1371/journal.pcbi.1000605>
- Schroeder, R., Barta, A., Semrad, K., 2004. Strategies for RNA folding and assembly. *Nat. Rev. Mol. Cell Biol.* 5, 908-919. <https://doi.org/10.1038/nrm1497>
- Schuler, B., Hofmann, H., 2013. Single-molecule spectroscopy of protein folding dynamics--expanding scope and timescales. *Curr. Opin. Struct. Biol.* 23, 36-47. <https://doi.org/10.1016/j.sbi.2012.10.008>
- Schütze, K., Heinzinger, M., Steinegger, M., Rost, B., 2022. Nearest neighbor search on embeddings rapidly identifies distant protein relations. <https://doi.org/10.1101/2022.09.04.506527>
- Sharma, R., Raduly, Z., Miskei, M., Fuxreiter, M., 2015. Fuzzy complexes: Specific binding without complete folding. *FEBS Lett.* 589, 2533-2542. <https://doi.org/10.1016/j.febslet.2015.07.022>
- Shimizu, K., Cao, W., Saad, G., Shoji, M., Terada, T., 2018. Comparative analysis of membrane protein structure databases. *Biochim. Biophys. Acta BBA - Biomembr.* 1860, 1077-1091. <https://doi.org/10.1016/j.bbamem.2018.01.005>
- Shindyalov, I.N., Bourne, P.E., 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng. Des. Sel.* 11, 739-747. <https://doi.org/10.1093/protein/11.9.739>
- Sillitoe, I., Bordin, N., Dawson, N., Waman, V.P., Ashford, P., Scholes, H.M., Pang, C.S.M., et al., 2021. CATH: increased structural coverage of functional space. *Nucleic Acids Res.* 49, D266-D273. <https://doi.org/10.1093/nar/gkaa1079>

- Sipiczki, M., 2011. Diversity, variability and fast adaptive evolution of the wine yeast (*Saccharomyces cerevisiae*) genome—a review. *Ann. Microbiol.* 61, 85-93. <https://doi.org/10.1007/s13213-010-0086-4>
- Sklenar, H., Etchebest, C., Lavery, R., 1989. Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins* 6, 46-60. <https://doi.org/10.1002/prot.340060105>
- Sommer, M.J., Cha, S., Varabyou, A., Rincon, N., Park, S., Minkin, I., Pertea, M., Steinegger, M., Salzberg, S.L., 2022. Highly accurate isoform identification for the human transcriptome. <https://doi.org/10.1101/2022.06.08.495354>
- Soria, P.S., McGary, K.L., Rokas, A., 2014. Functional Divergence for Every Paralog. *Mol. Biol. Evol.* 31, 984-992. <https://doi.org/10.1093/molbev/msu050>
- Soyer, A., Chomilier, J., Morion, J.-P., Jullien, R., Sadoc, J.-F., 2000. Voronoï Tessellation Reveals the Condensed Matter Character of Folded Proteins. *Phys. Rev. Lett.* 85, 3532-5. <https://doi.org/10.1103/PhysRevLett.85.3532>
- Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., Söding, J., 2019a. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* 20, 473. <https://doi.org/10.1186/s12859-019-3019-7>
- Steinegger, M., Mirdita, M., Söding, J., 2019b. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods* 16, 603-606. <https://doi.org/10.1038/s41592-019-0437-4>
- Steinegger, M., Söding, J., 2018. Clustering huge protein sequence sets in linear time. *Nat. Commun.* 9, 2542. <https://doi.org/10.1038/s41467-018-04964-5>
- Storz, G., Wolf, Y.I., Ramamurthi, K.S., 2014. Small proteins can no longer be ignored. *Annu. Rev. Biochem.* 83, 753-777. <https://doi.org/10.1146/annurev-biochem-070611-102400>
- Sugase, K., Dyson, H.J., Wright, P.E., 2007. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* 447, 1021-1025. <https://doi.org/10.1038/nature05858>
- Sun, Y., Shang, L., Zhu, Q.-H., Fan, L., Guo, L., 2022. Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci.* 27, 391-401. <https://doi.org/10.1016/j.tplants.2021.10.006>
- Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., et al., 2015. Structure and function of the global ocean microbiome. *Science* 348, 1261359. <https://doi.org/10.1126/science.1261359>
- Tautz, D., Domazet-Lošo, T., 2011. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 12, 692-702. <https://doi.org/10.1038/nrg3053>
- Taylor, W.R., 1986. The classification of amino acid conservation. *J. Theor. Biol.* 119, 205-218. [https://doi.org/10.1016/S0022-5193\(86\)80075-3](https://doi.org/10.1016/S0022-5193(86)80075-3)
- Teraguchi, S., Patil, A., Standley, D.M., 2010. Intrinsically disordered domains deviate significantly from random sequences in mammalian proteins. *BMC Bioinformatics* 11, S7. <https://doi.org/10.1186/1471-2105-11-S7-S7>
- The UniProt Consortium, 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480-D489. <https://doi.org/10.1093/nar/gkaa1100>
- Then, A., Mácha, K., Ibrahim, B., Schuster, S., 2020. A novel method for achieving an optimal classification of the proteinogenic amino acids. *Sci. Rep.* 10, 15321. <https://doi.org/10.1038/s41598-020-72174-5>
- Thomas, A.M., Segata, N., 2019. Multiple levels of the unknown in microbiome research. *BMC Biol.* 17, 48. <https://doi.org/10.1186/s12915-019-0667-z>
- Thornton, J.M., Laskowski, R.A., Borkakoti, N., 2021. AlphaFold heralds a data-driven revolution in biology and medicine. *Nat. Med.* 27, 1666-1669. <https://doi.org/10.1038/s41591-021-01533-0>
- Tompa, P., 2005. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.* 579, 3346-3354. <https://doi.org/10.1016/j.febslet.2005.03.072>
- Tompa, P., 2002. Intrinsically unstructured proteins. *Trends Biochem. Sci.* 27, 527-533. [https://doi.org/10.1016/s0968-0004\(02\)02169-2](https://doi.org/10.1016/s0968-0004(02)02169-2)
- Tompa, P., Fuxreiter, M., 2008. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.* 33, 2-8. <https://doi.org/10.1016/j.tibs.2007.10.003>

- Tompa, P., Fuxreiter, M., Oldfield, C.J., Simon, I., Dunker, A.K., Uversky, V.N., 2009. Close encounters of the third kind: disordered domains and the interactions of proteins. *BioEssays News Rev. Mol. Cell. Dev. Biol.* 31, 328-335. <https://doi.org/10.1002/bies.200800151>
- Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Židek, A., Bridgland, A., et al., 2021. Highly accurate protein structure prediction for the human proteome. *Nature* 596, 590-596. <https://doi.org/10.1038/s41586-021-03828-1>
- Tusnády, G.E., Dosztányi, Z., Simon, I., 2004. Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics* 20, 2964-2972. <https://doi.org/10.1093/bioinformatics/bth340>
- Uversky, V.N., Dunker, A.K., 2010. Understanding protein non-folding. *Biochim. Biophys. Acta* 1804, 1231-1264. <https://doi.org/10.1016/j.bbapap.2010.01.017>
- Uversky, V.N., Gillespie, J.R., Fink, A.L., 2000. Why are « natively unfolded » proteins unstructured under physiologic conditions? *Proteins* 41, 415-427. [https://doi.org/10.1002/1097-0134\(20001115\)41:3<415::aid-prot130>3.0.co;2-7](https://doi.org/10.1002/1097-0134(20001115)41:3<415::aid-prot130>3.0.co;2-7)
- Vacic, V., Oldfield, C.J., Mohan, A., Radivojac, P., Cortese, M.S., Uversky, V.N., Dunker, A.K., 2007. Characterization of Molecular Recognition Features, MoRFs, and Their Binding Partners. *J. Proteome Res.* 6, 2351-2366. <https://doi.org/10.1021/pr0701411>
- Vakirlis, N., Acar, O., Hsu, B., Castilho Coelho, N., Van Oss, S.B., Wacholder, A., Medetgul-Ernar, K., et al., 2020. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat. Commun.* 11, 781. <https://doi.org/10.1038/s41467-020-14500-z>
- Vakirlis, N., Hebert, A.S., Opulente, D.A., Achaz, G., Hittinger, C.T., Fischer, G., Coon, J.J., Lafontaine, I., 2018. A Molecular Portrait of De Novo Genes in Yeasts. *Mol. Biol. Evol.* 35, 631-645. <https://doi.org/10.1093/molbev/msx315>
- van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., et al., 2014. Classification of intrinsically disordered regions and proteins. *Chem. Rev.* 114, 6589-6631. <https://doi.org/10.1021/cr400525m>
- Vanni, C., Schechter, M.S., Acinas, S.G., Barberán, A., Buttigieg, P.L., Casamayor, E.O., Delmont, T.O., et al., 2022. Unifying the known and unknown microbial coding sequence space. *eLife* 11, e67667. <https://doi.org/10.7554/eLife.67667>
- Vanni, C., Schechter, M.S., Delmont, T.O., Eren, A.M., Steinegger, M., Glöckner, F.O., Fernandez-Guerra, A., 2021. AGNOSTOS-DB: a resource to unlock the uncharted regions of the coding sequence space. <https://doi.org/10.1101/2021.06.07.447314>
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., et al., 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50, D439-D444. <https://doi.org/10.1093/nar/gkab1061>
- Wang, J., Choi, J.-M., Holehouse, A.S., Lee, H.O., Zhang, X., Jahnke, M., Maharana, S., et al., 2018. A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell* 174, 688-699.e16. <https://doi.org/10.1016/j.cell.2018.06.006>
- Wang, L., Wang, M., Shi, X., Yang, J., Qian, C., Liu, Q., Zong, L., et al., 2020. Investigation into archaeal extremophilic lifestyles through comparative proteogenomic analysis. *J. Biomol. Struct. Dyn.* 0, 1-13. <https://doi.org/10.1080/07391102.2020.1808531>
- Wang, S., Ma, J., Xu, J., 2016. AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. *Bioinformatics* 32, i672-i679. <https://doi.org/10.1093/bioinformatics/btw446>
- Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., Jones, D.T., 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* 337, 635-645. <https://doi.org/10.1016/j.jmb.2004.02.002>
- Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., et al., 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296-W303. <https://doi.org/10.1093/nar/gky427>
- Weatheritt, R.J., Luck, K., Petsalaki, E., Davey, N.E., Gibson, T.J., 2012. The identification of short linear motif-mediated interfaces within the human interactome. *Bioinformatics* 28, 976-982. <https://doi.org/10.1093/bioinformatics/bts072>

- Webb, B., Sali, A., 2021. Protein Structure Modeling with MODELLER, in: Chen, Y.W., Yiu, C.-P.B. (Éd.), *Structural Genomics: General Applications, Methods in Molecular Biology*. Springer US, New York, NY, p. 239-255. [https://doi.org/10.1007/978-1-0716-0892-0\\_14](https://doi.org/10.1007/978-1-0716-0892-0_14)
- Weinbauer, M.G., Rassoulzadegan, F., 2004. Are viruses driving microbial diversification and diversity? *Environ. Microbiol.* 6, 1-11. <https://doi.org/10.1046/j.1462-2920.2003.00539.x>
- Weisman, C.M., Murray, A.W., Eddy, S.R., 2020. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLOS Biol.* 18, e3000862. <https://doi.org/10.1371/journal.pbio.3000862>
- Weissenow, K., Heinzinger, M., Rost, B., 2022. Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure* 30, 1169-1177.e4. <https://doi.org/10.1016/j.str.2022.05.001>
- Wetlaufer, D.B., 1973. Protein Structure and Stability: Conventional Wisdom and New Perspectives. *J. Food Sci.* 38, 740-743. <https://doi.org/10.1111/j.1365-2621.1973.tb02068.x>
- Wheelan, S.J., Marchler-Bauer, A., Bryant, S.H., 2000. Domain size distributions can predict domain boundaries. *Bioinformatics* 16, 613-618. <https://doi.org/10.1093/bioinformatics/16.7.613>
- White, S.H., 2009. Biophysical dissection of membrane proteins. *Nature* 459, 344-346. <https://doi.org/10.1038/nature08142>
- Wilson, B.A., Masel, J., 2011. Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol. Evol.* 3, 1245-1252. <https://doi.org/10.1093/gbe/evr099>
- Wilson, C.J., Choy, W.-Y., Karttunen, M., 2021. AlphaFold2: A role for disordered protein prediction? <https://doi.org/10.1101/2021.09.27.461910>
- Wolf, Y.I., Grishin, N.V., Koonin, E.V., 2000. Estimating the number of protein folds and families from complete genome data. Edited by J. Thornton. *J. Mol. Biol.* 299, 897-905. <https://doi.org/10.1006/jmbi.2000.3786>
- Woodcock, S., Mornon, J.P., Henrissat, B., 1992. Detection of secondary structure elements in proteins by hydrophobic cluster analysis. *Protein Eng.* 5, 629-635. <https://doi.org/10.1093/protein/5.7.629>
- Wootton, J.C., Federhen, S., 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* 17, 149-163. [https://doi.org/10.1016/0097-8485\(93\)85006-X](https://doi.org/10.1016/0097-8485(93)85006-X)
- Wright, P.E., Dyson, H.J., 2015. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* 16, 18-29. <https://doi.org/10.1038/nrm3920>
- Wright, P.E., Dyson, H.J., 2009. Linking folding and binding. *Curr. Opin. Struct. Biol.* 19, 31-38. <https://doi.org/10.1016/j.sbi.2008.12.003>
- Wright, P.E., Dyson, H.J., 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* 293, 321-331. <https://doi.org/10.1006/jmbi.1999.3110>
- Wu, B., Knudson, A., 2018. Tracing the De Novo Origin of Protein-Coding Genes in Yeast. *mBio* 9. <https://doi.org/10.1128/mBio.01024-18>
- Wu, H., Fuxreiter, M., 2016. The Structure and Dynamics of Higher-Order Assemblies: Amyloids, Signalosomes, and Granules. *Cell* 165, 1055-1066. <https://doi.org/10.1016/j.cell.2016.05.004>
- Wu, Y.-C., Rasmussen, M.D., Kellis, M., 2012. Evolution at the Subgene Level: Domain Rearrangements in the Drosophila Phylogeny. *Mol. Biol. Evol.* 29, 689-705. <https://doi.org/10.1093/molbev/msr222>
- Wyman, S.K., Avila-Herrera, A., Nayfach, S., Pollard, K.S., 2018. A most wanted list of conserved microbial protein families with no known domains. *PloS One* 13, e0205749. <https://doi.org/10.1371/journal.pone.0205749>
- Xue, B., Williams, R.W., Oldfield, C.J., Dunker, A.K., Uversky, V.N., 2010. Archaic chaos: intrinsically disordered proteins in Archaea. *BMC Syst. Biol.* 4, S1. <https://doi.org/10.1186/1752-0509-4-S1-S1>
- Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J.A., et al., 2007. The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. *PLOS Biol.* 5, e16. <https://doi.org/10.1371/journal.pbio.0050016>
- Young, J.C., Agashe, V.R., Siegers, K., Hartl, F.U., 2004. Pathways of chaperone-mediated protein folding in the cytosol. *Nat. Rev. Mol. Cell Biol.* 5, 781-791. <https://doi.org/10.1038/nrm1492>

- Zarin, T., Strome, B., Nguyen Ba, A.N., Alberti, S., Forman-Kay, J.D., Moses, A.M., 2019. Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *eLife* 8, e46883. <https://doi.org/10.7554/eLife.46883>
- Zhang, C., DeLisi, C., 1998. Estimating the number of protein folds. Edited by F. Cohen. *J. Mol. Biol.* 284, 1301-1305. <https://doi.org/10.1006/jmbi.1998.2282>
- Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A.R., Yu, Y., et al., 2019. Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nat. Ecol. Evol.* <https://doi.org/10.1038/s41559-019-0822-5>
- Zhang, Tuo, Faraggi, E., Xue, B., Dunker, A.K., Uversky, V.N., Zhou, Y., 2012. SPINE-D: Accurate Prediction of Short and Long Disordered Regions by a Single Neural-Network Based Method. *J. Biomol. Struct. Dyn.* 29, 799-813. <https://doi.org/10.1080/073911012010525022>
- Zhang, Tongwu, Hu, S., Zhang, G., Pan, L., Zhang, X., Al-Mssallem, I.S., Yu, J., 2012. The Organelle Genomes of Hassawi Rice (*Oryza sativa* L.) and Its Hybrid in Saudi Arabia: Genome Variation, Rearrangement, and Origins. *PLOS ONE* 7, e42041. <https://doi.org/10.1371/journal.pone.0042041>
- Zhang, Y., Skolnick, J., 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* 33, 2302-2309. <https://doi.org/10.1093/nar/gki524>

## RESUME

Une fraction significative des protéomes reste non annotée, laissant inaccessible une partie du répertoire fonctionnel de la vie, incluant des innovations moléculaires ayant une valeur thérapeutique ou environnementale. Le manque d'annotation fonctionnelle est en partie dû aux limites des approches actuelles pour la détection de relations cachées, ou à des caractéristiques spécifiques telles que le désordre.

L'objectif de ma thèse a été de développer des approches méthodologiques reposant sur les signatures structurales des domaines repliés, afin de caractériser plus avant les séquences protéiques dont la fonction est inconnue, même en l'absence d'informations évolutives.

Tout d'abord, j'ai développé un score permettant d'estimer le potentiel de repliement d'une séquence d'acides aminés, basé sur sa densité en amas hydrophobes, correspondant principalement aux structures secondaires régulières. J'ai décrit le continuum entre l'ordre et le désordre, couvrant différents états allant des conformations étendues aux globules fondus et ai caractérisé des cas d'ordre conditionnel. Ensuite, j'ai combiné ce score avec les prédictions de structure 3D d'AlphaFold2 (AF2) disponibles pour 21 protéomes de référence. Une grande fraction des acides aminés des modèles AF2 associés à un très faible index de confiance est incluse dans des segments non repliables, soutenant la qualité d'AF2 comme prédicteur du désordre. Cependant, dans chaque protéome, de longs segments repliables avec des prédictions AF2 de faible confiance présentent également des caractéristiques de domaines solubles et repliés. Cela suggère un ordre caché (conditionnel ou inconditionnel), qui n'est pas détecté par AF2 en raison du manque d'informations évolutives, ou des motifs de repliement non répertoriés. Enfin, à l'aide de ces outils, j'ai effectué une exploration préliminaire de protéines ou de régions non annotées, identifiées via le développement et l'application d'une nouvelle procédure d'annotation. Bien que ces séquences soient enrichies en désordre, une part importante d'entre elles présente des caractéristiques de type globulaire soluble. Ces séquences constituent de bons candidats pour de futures validations et caractérisations expérimentales. De plus, l'analyse de gènes *de novo* validés expérimentalement m'a permis de contribuer au débat encore ouvert sur les caractéristiques structurales des protéines codées par ces gènes, qui présentent un enrichissement en désordre et une grande diversité d'états structuraux.

**Mots-clés** : dark protéome, amas hydrophobes, segments repliables, protéines et régions intrinsèquement désordonnées, ordre caché, AlphaFold2

## Leveraging protein fold signatures to describe the order/disorder continuum in proteomes

### ABSTRACT

A significant fraction of the proteomes remains unannotated, leaving inaccessible a part of the functional repertoire of life, including molecular innovations with therapeutic or environmental value. Lack of functional annotation is partly due to the limitations of the current approaches in detecting hidden relationships, or to specific features such as disorder. The aim of my PhD thesis was to develop methodological approaches based on the structural signatures of folded domains, in order to further characterize the protein sequences with unknown function even in absence of evolutionary information.

First, I developed a scoring system in order to estimate the foldability potential of an amino acid sequence, based on its density in hydrophobic clusters, which mainly correspond to regular secondary structures. I disentangled the continuum between order and disorder, covering various states from extended conformations (random coils) to molten globules and characterize cases of conditional order. Next, I combined this scoring system with the AlphaFold2 (AF2) 3D structure predictions available for 21 reference proteomes. A large fraction of the amino acids with very low AF2 model confidence are included in non-foldable segments, supporting the quality of AF2 as a predictor of disorder. However, within each proteome, long segments with very low AF2 model confidence also exhibit characteristics of soluble, folded domains. This suggests hidden order (conditional or unconditional), which is undetected by AF2 due to lack of evolutionary information, or unrecorded folding patterns. Finally, using these tools, I made a preliminary exploration of unannotated proteins or regions, identified through the development and application of a new annotation workflow. Even though these sequences are enriched in disorder, an important part of them showcases soluble globular-like characteristics. These would make good candidates for further experimental validation and characterization. Moreover, the analysis of experimentally validated *de novo* genes allowed me to contribute to the still-open debate on the structural features of proteins encoded by these genes, enriched in disorder and displaying a great diversity of structural states.

**Keywords**: dark proteome, hydrophobic clusters, foldable segments, IDPs/IDRs, hidden order, AlphaFold2