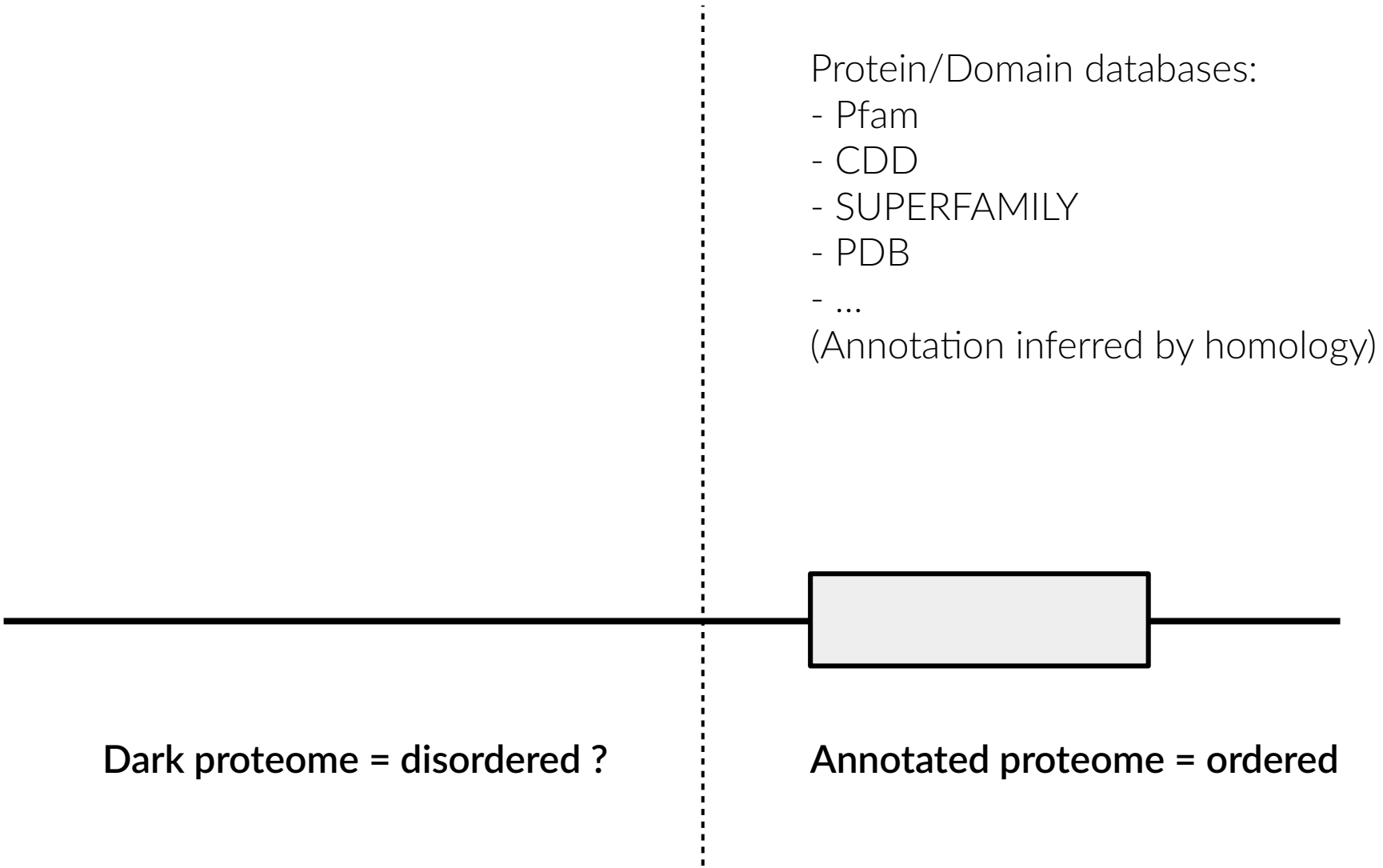
Looking in the dark:
disorder associated with “unannotated” protein sequences and emergence of de novo protein domains

Tristan Bitard-Feildel

Laboratoire de Biologie Computationnelle et Quantitative

(Institut for Evolution and Biodiversity &
Institut de Minéralogie de Physique de la Matière et de Cosmochimie)

Introduction



Introduction

Sequence divergence ?
Species bias ?
Amino acid compositional bias ?
Missing family ?

Protein/Domain databases:
- Pfam
- CDD
- SUPERFAMILY
- PDB
- ...

(Annotation inferred by homology)

Dark proteome = disordered ?

Annotated proteome = ordered

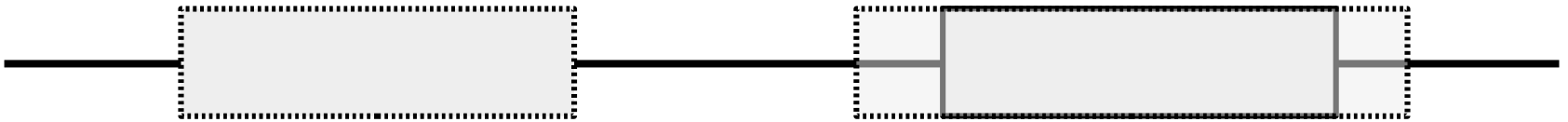
Introduction

HCA based delineation → no homology

- Physicochemical properties (IDR?)
- Evolutionary events
- Tools

Protein/Domain databases:
- Pfam
- CDD
- SUPERFAMILY
- PDB
- ...

(Annotation inferred by homology)



HCA segment, orphan domain (orphan of any annotation)

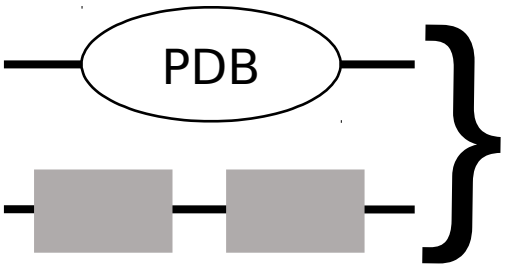
Ordered

Uniprot/Swissprot dataset

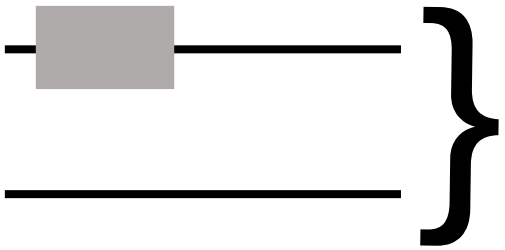
> 550 000 sequences (curated)

PDBe
Protein Data Bank in Europe

Several domain databases (HMM, PSSM, structures)

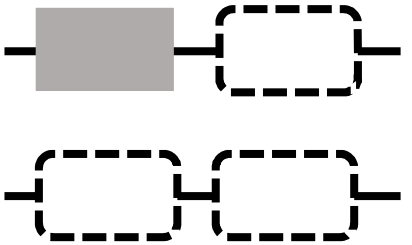


Structural or domain annotation



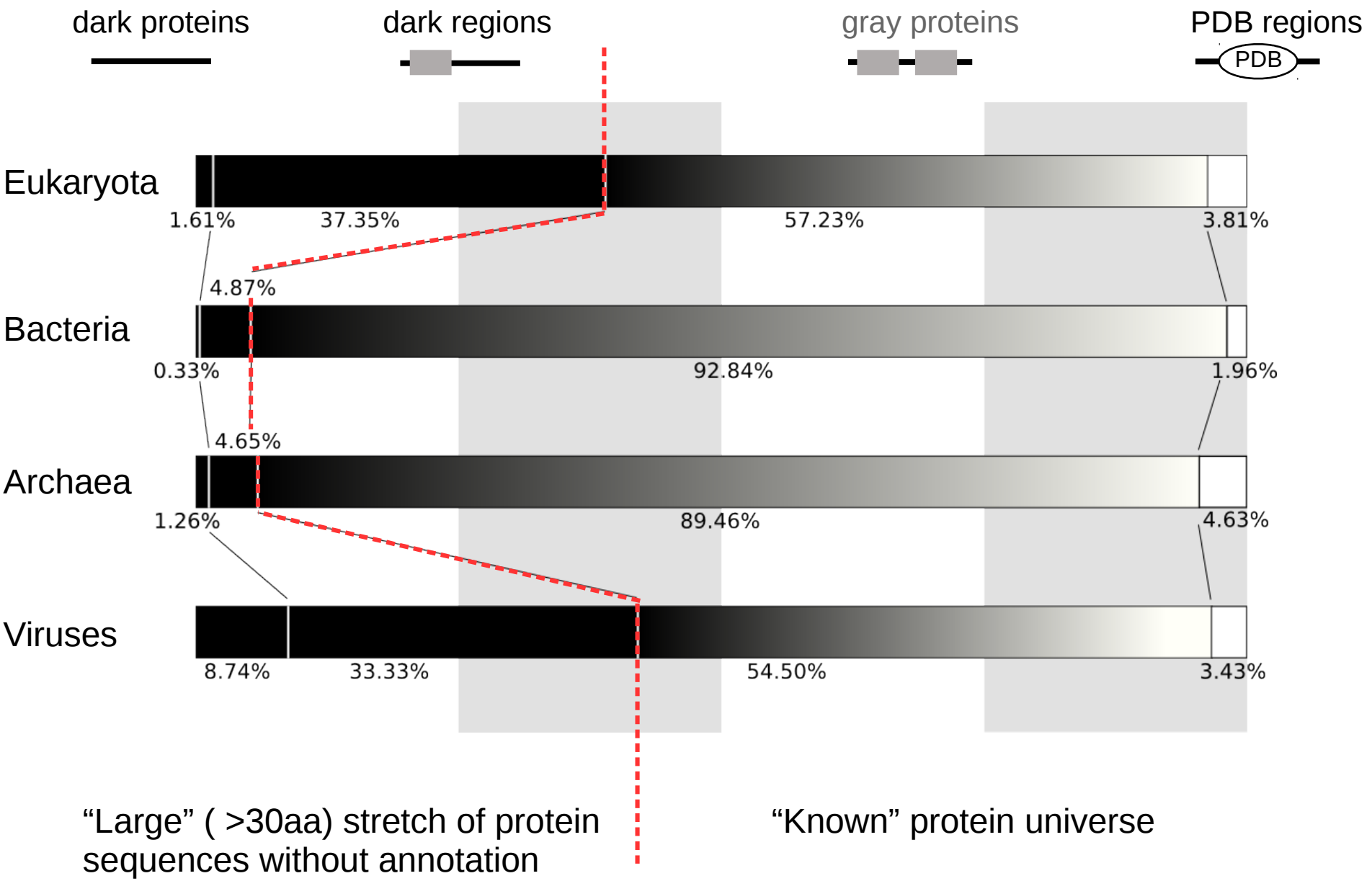
Part or whole protein without annotation

Hydrophobic Cluster Analysis



T. Bitard-Feildel, I. Callebaut – Scientific reports 2017

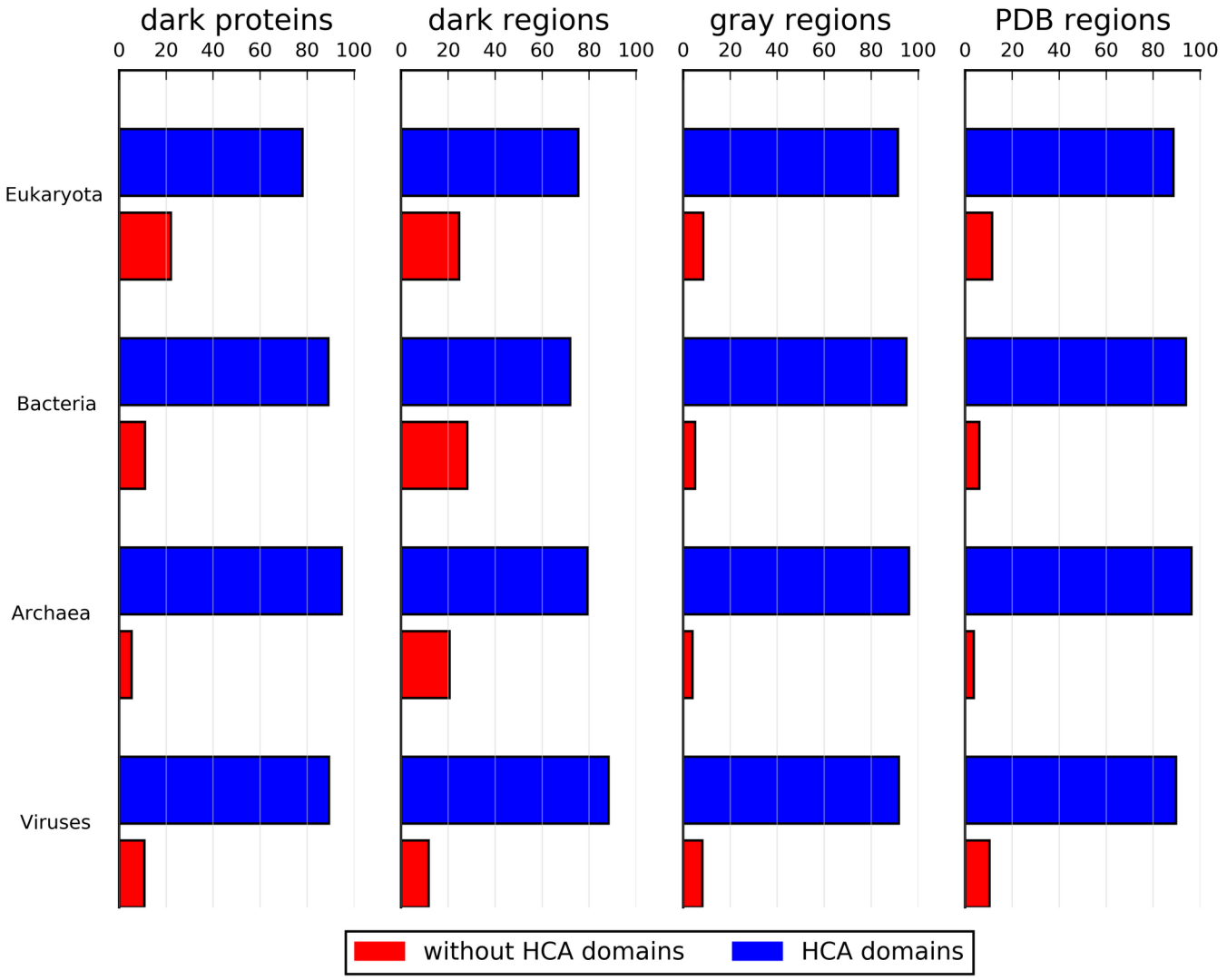
Uniprot/Swissprot dataset



“Large” (>30aa) stretch of protein sequences without annotation

“Known” protein universe

Uniprot/Swissprot dataset

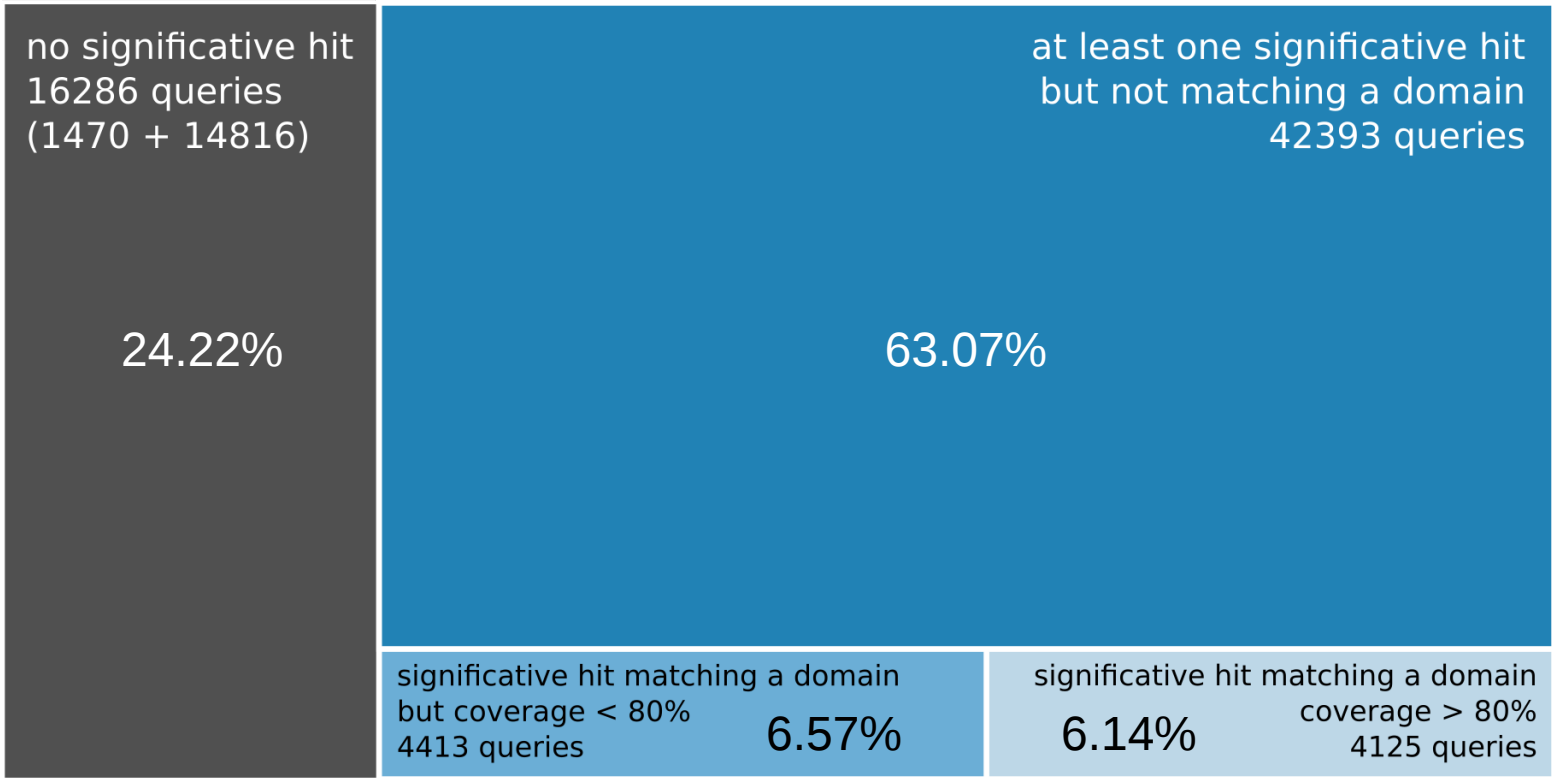


80% to 90% of protein sequences correspond to an HCA domains

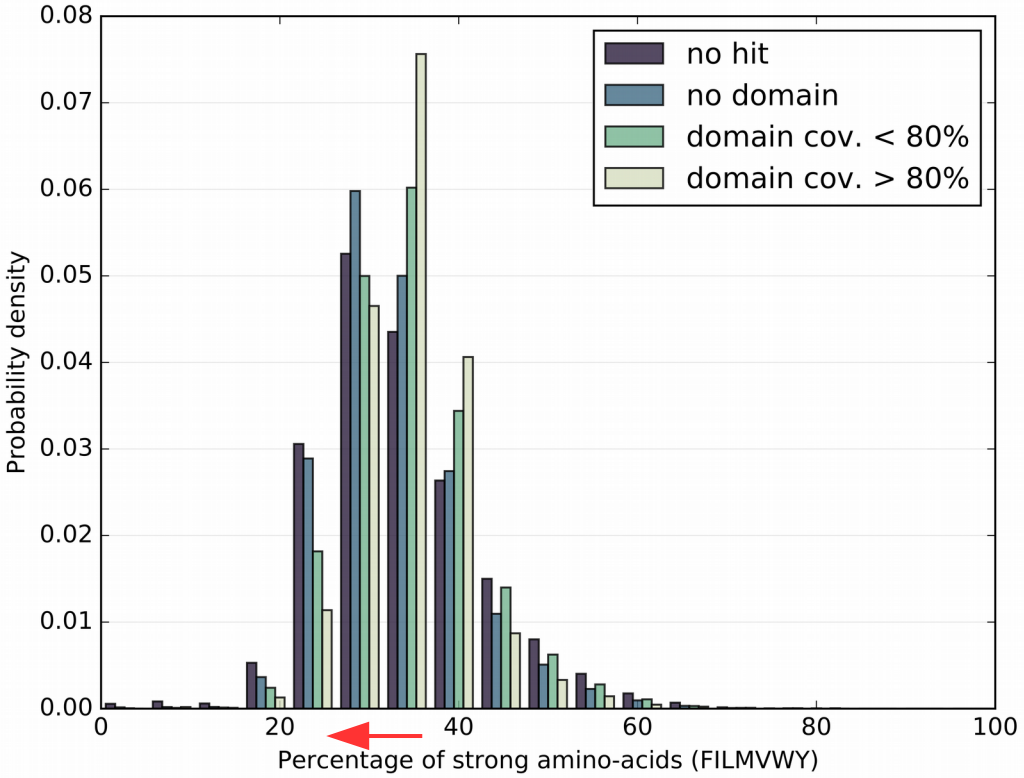
Uniprot/Swissprot dataset

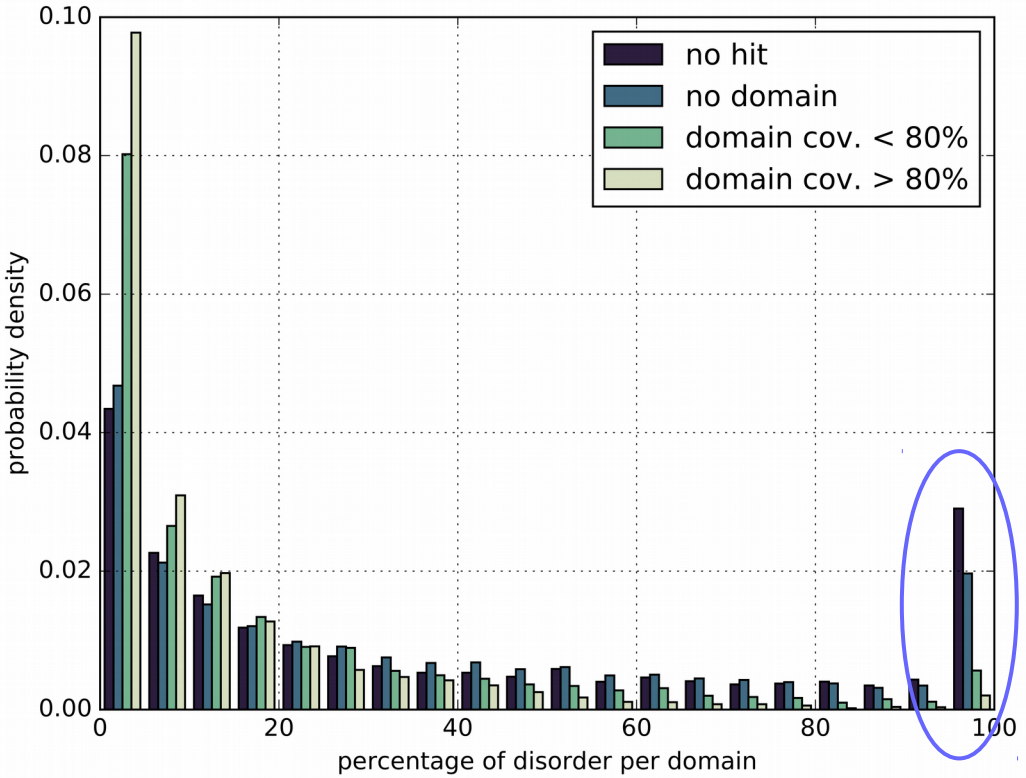
After filtering/clustering: 67 217 dark domains

For each domain, use of Tremolo-HCA (Hhblits+HCA+phylogeny) to find remote similarity



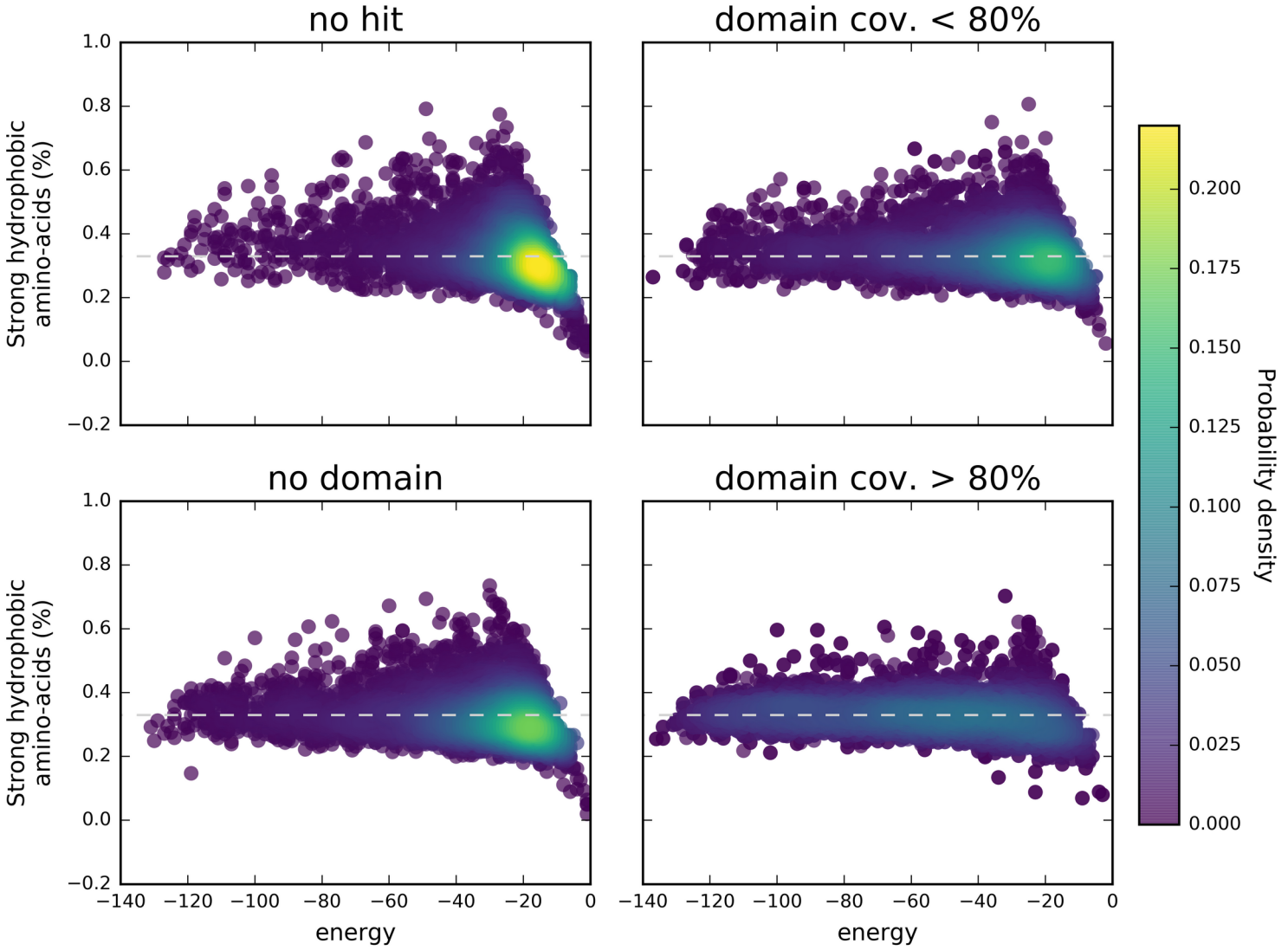
Uniprot/Swissprot dataset

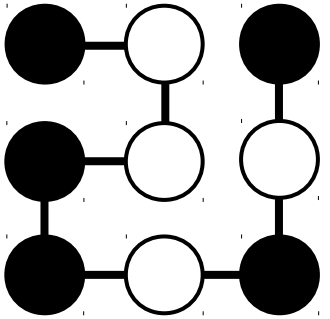




Fewer strong amino acids and higher disorder content
for sequences with without homology and “annotation”

Uniprot/Swissprot dataset





Lower “folding energy” for domain sequences without “annotation”

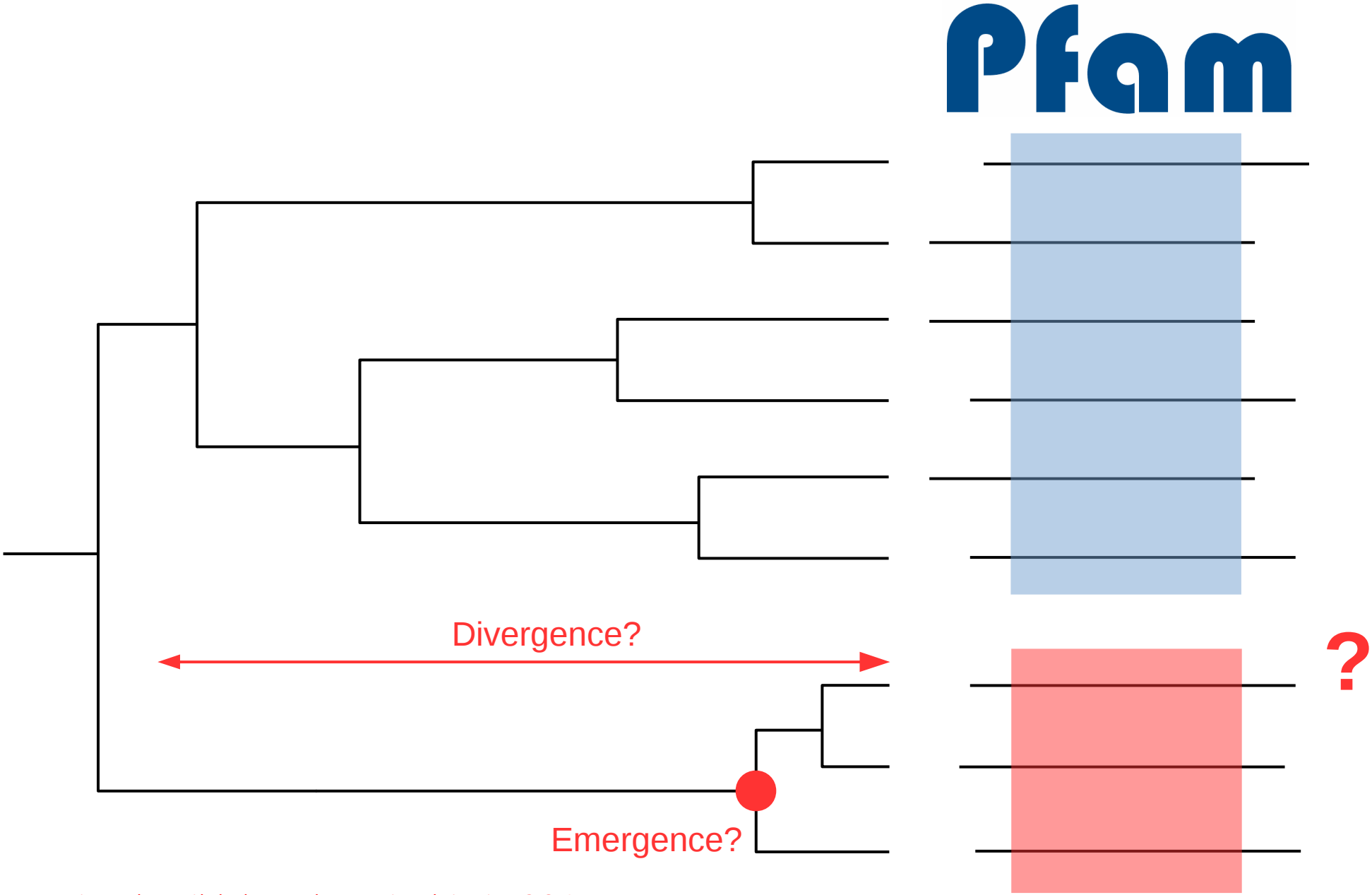
Sequence of the **dark proteome**:
- are mostly Eukaryotic and Viral protein sequences
- can be **delineated in segments/domains** based on their hydrophobicity

HCA segments of the dark proteome:
- can be associated to known protein domains (a few)
- are without homologous proteins (many)

Dark HCA segments have:
- **fewer strong hydrophobic amino acids**
- **higher predicted disorder**
- **higher folding energy**

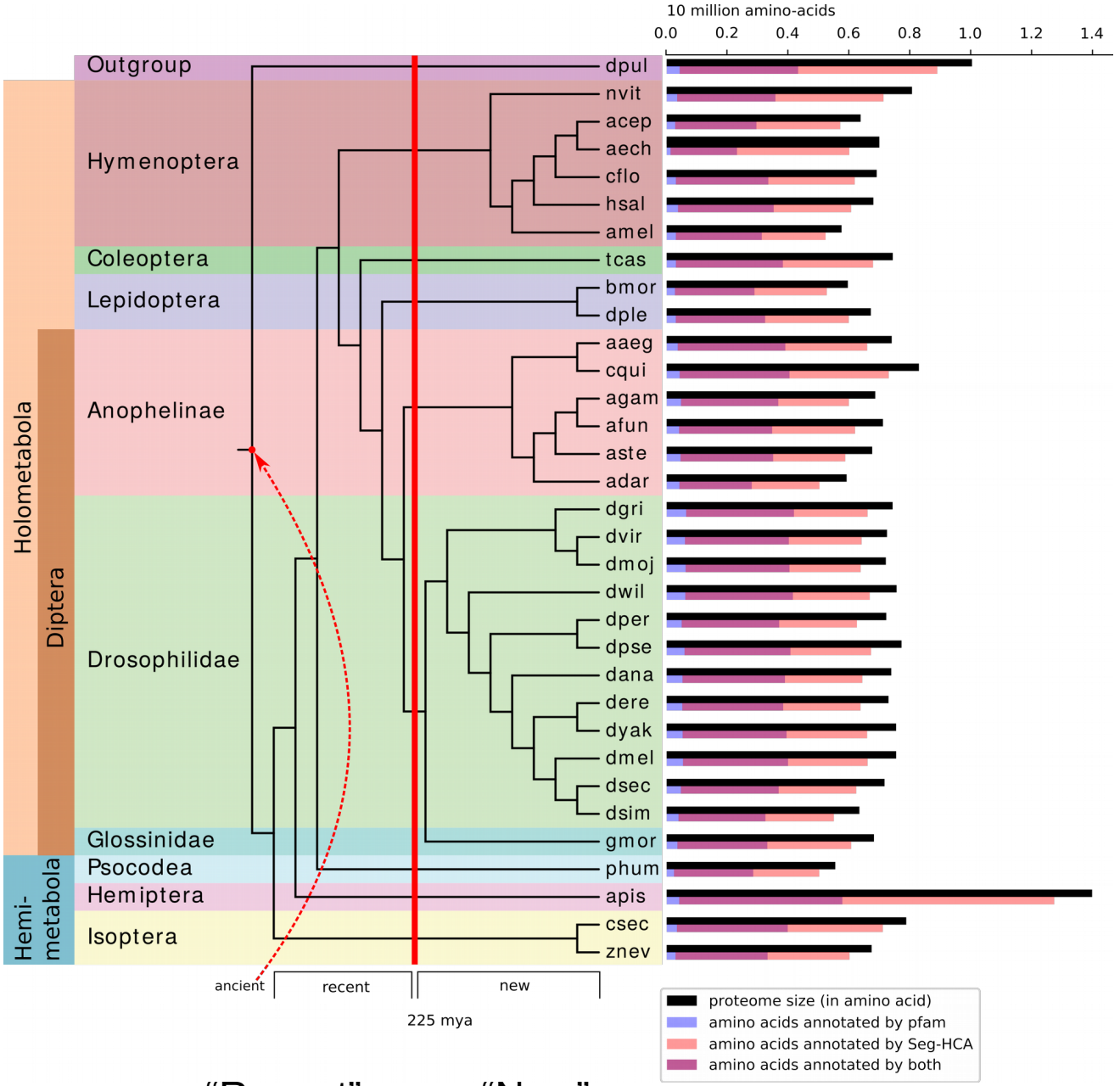
than HCA segments related to known domains

De novo protein domains



T. Bitard-Feildel et al – Biochimie 2015
T. Bitard-Feildel et al – FEBS J. 2018

De novo protein domains



“Recent” “New”

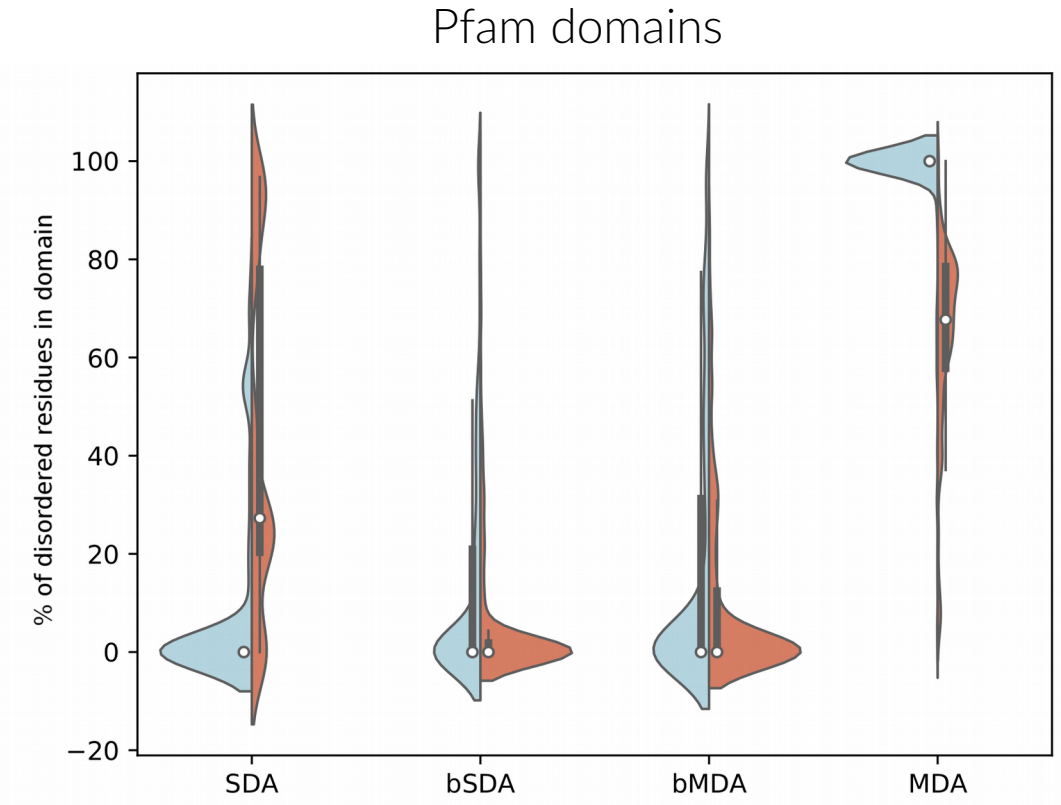
De novo protein domains

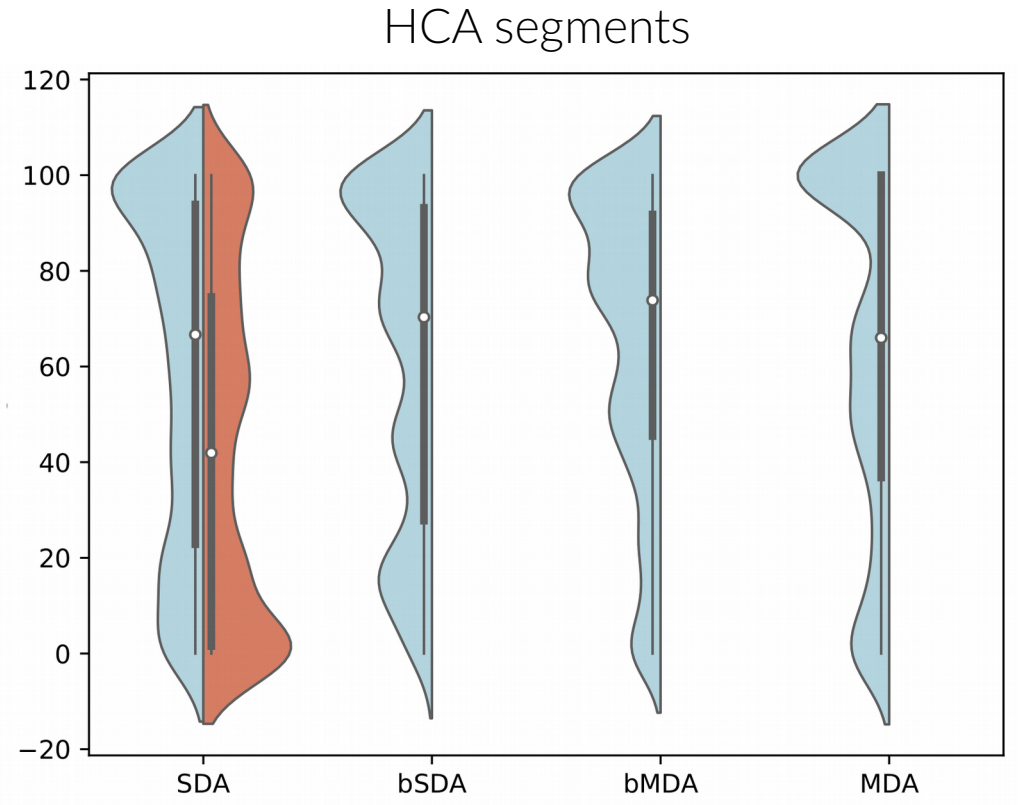
Pfam → 50 domain families taxonomically restricted

HCA → 177 putative domain families taxo. restricted

Spanning different emergence time

De novo protein domains



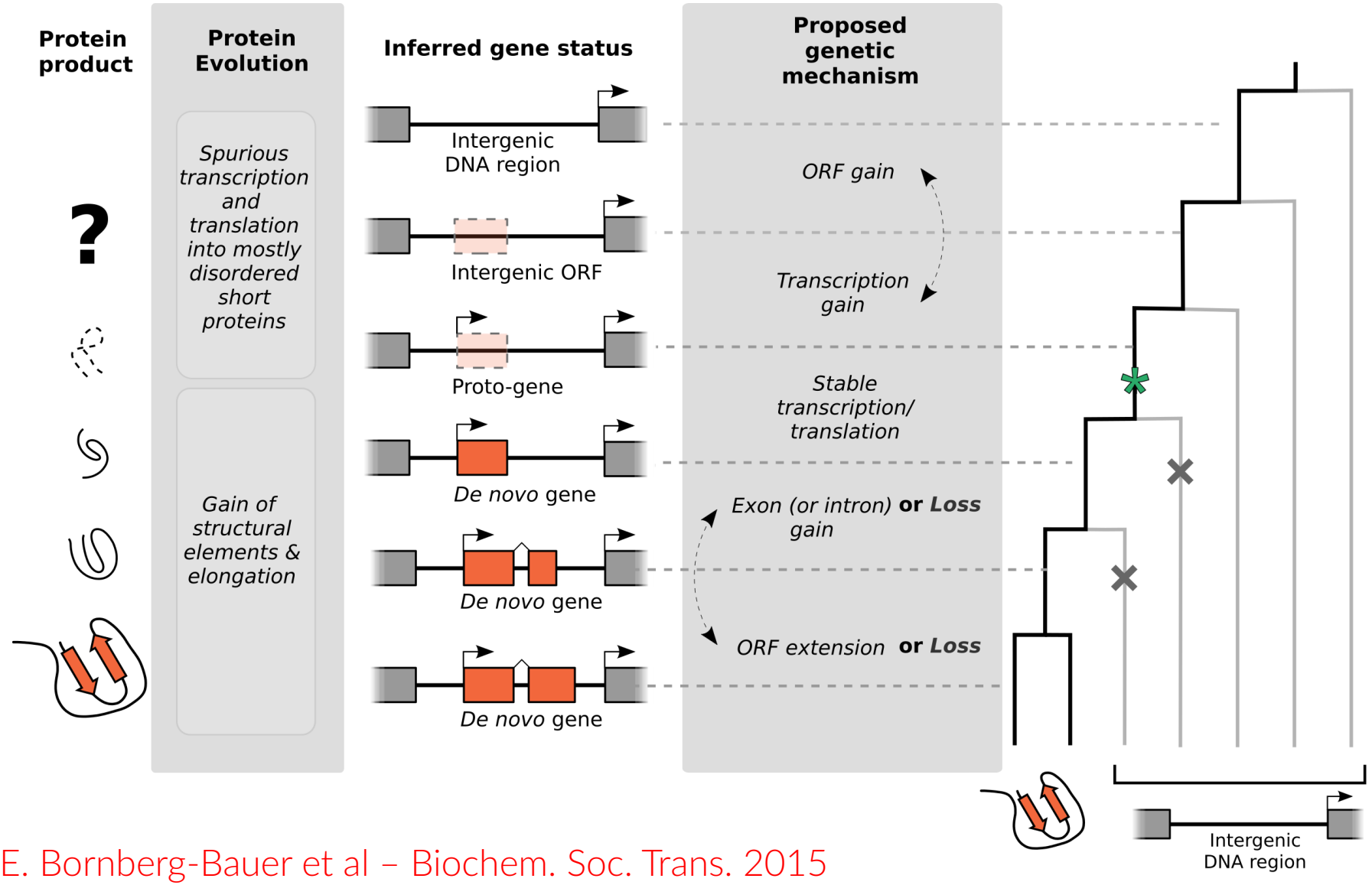


- **higher disorder** content in HCA segments than in Pfam domains
- orphan Pfam multidomain arrangements (MDA) with higher disorder content than other Pfam arrangements

- Pfam domains mostly as **C-termini** extension
- HCA domains mostly as **N-termini and middle positions**
- **Under less purifying selection** than “ancient domains”

De novo protein domains

Similar to some novel gene emergence model (growth slow and moult model)



E. Bornberg-Bauer et al – Biochem. Soc. Trans. 2015
JF Schmitz and E. Bornberg-Bauer – F1000 Research 2017

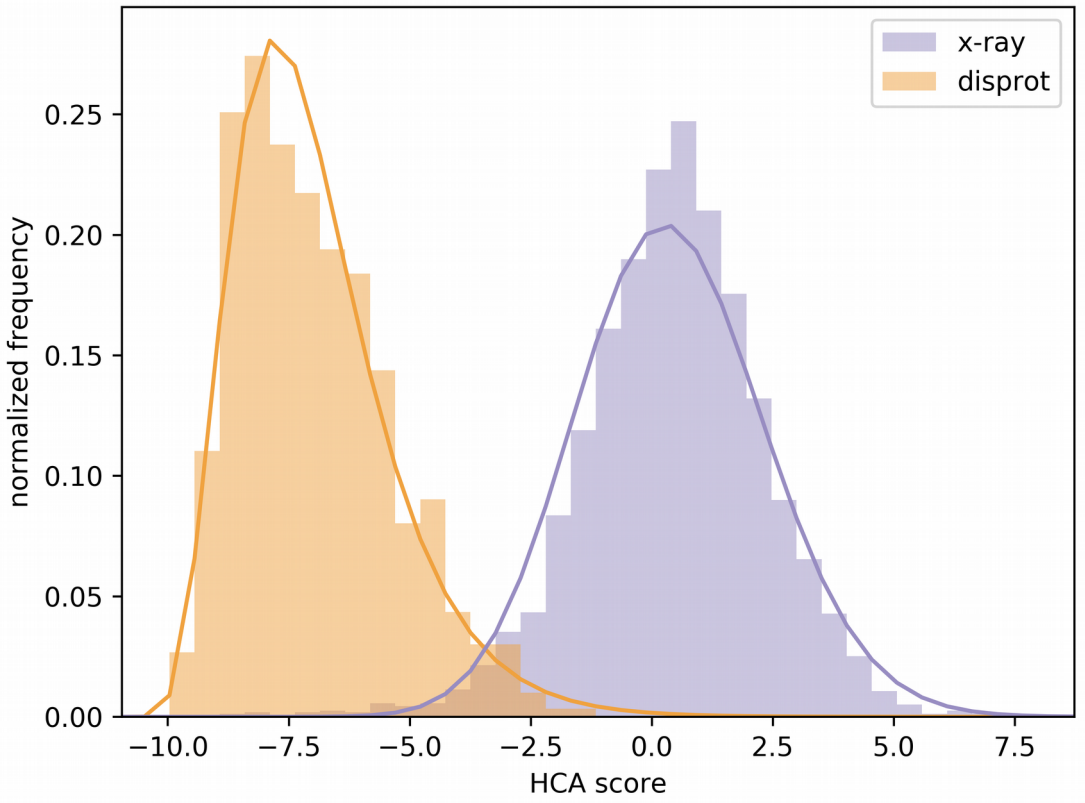
Hydrophobic Cluster Analysis

IDRs/IDPs (DisProt) VS Globular proteins (PDB)

MSGNQ**MAM**GSEQQQTVGSRT**VSVEEV**PA**VLQL**RATQDPPRSQEAMP

Number of hydrophobic and hydrophilic residues in clusters:
10000**101**0000000100000**10100**1**00110**1000000000000010

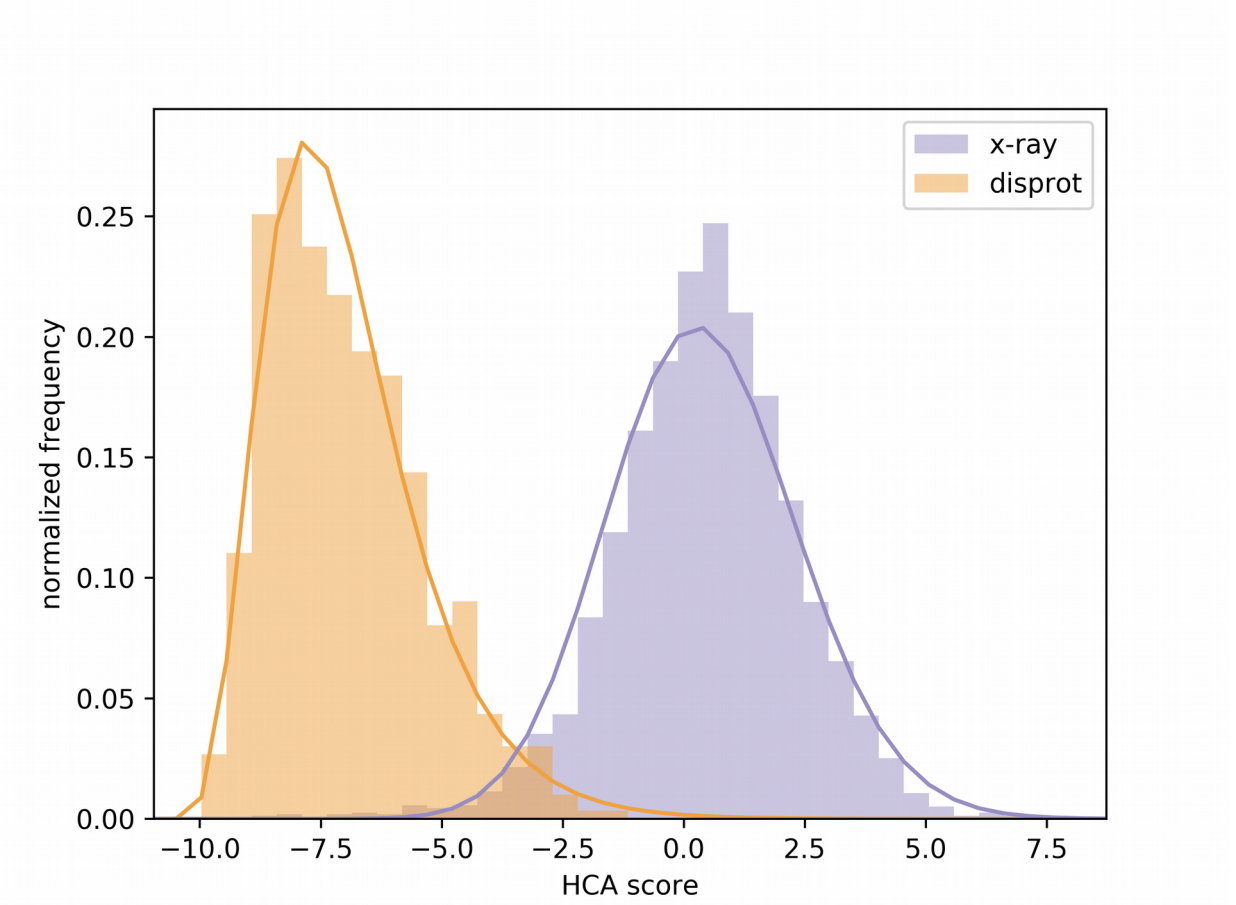
Number of residues not in clusters:
10000101**0000000100000**10100**00**1101**0000000000010**



Hydrophobic Cluster Analysis

IDRs/IDPs (DisProt) VS Globular proteins (PDB)

- Score at the HCA segment level
- P-value to determine the “foldability” of a segment



Hydrophobic Cluster Analysis

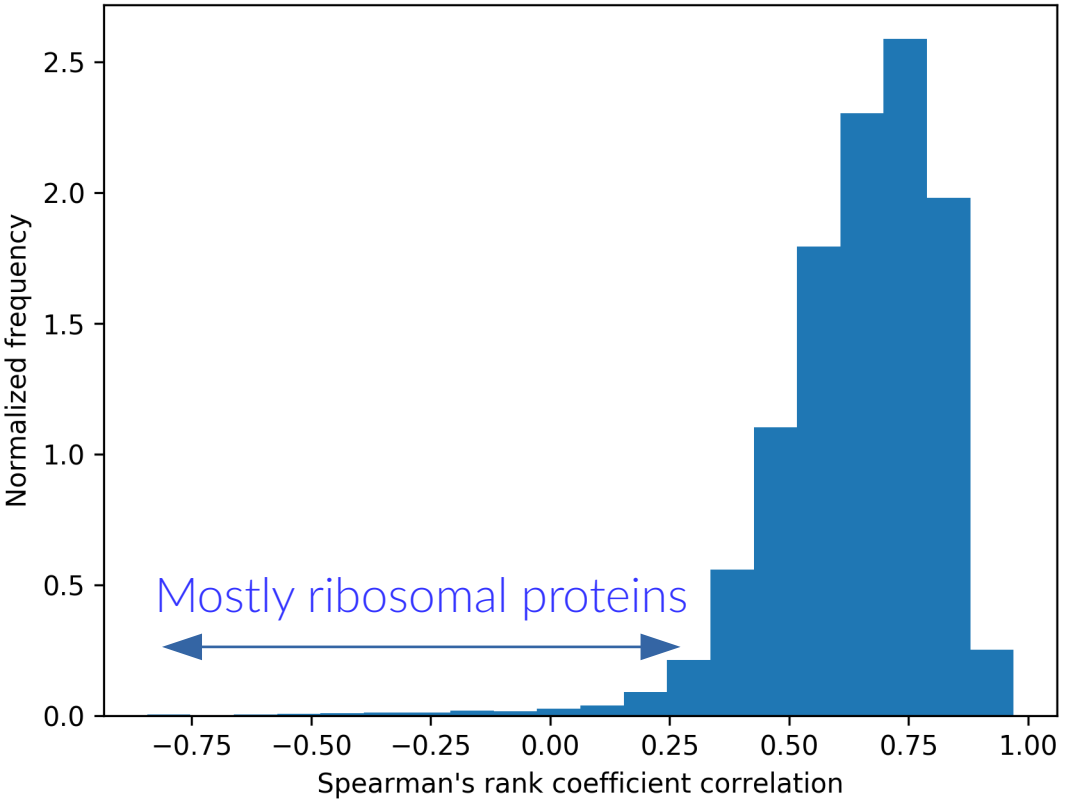
Residue based score

MSGNQ**MAM**GSEQQQTVGSRT**VSVEEV**PA**VLQL**RATQDPPRSQEAMP

1000**101**0000000010000**101001**00**1101**0000000000010

P(res. disorder) ~ Logistic regression classifier

HCA disorder score and Mobidb-Lite prediction correlation (protein level)



Hydrophobic Cluster Analysis

HCAtk:
- standalone program
- **segments** protein sequences
- search remote homology using **custom profiles**
- draw HCA diagrams
- predict **disorder** at the domain and residue level

pyHCA:
- a python API
- all the previous functionalities can be reused and interfaced in other python programs

Works with amino acids and nucleotidiques sequences (search on six frames)

https://github.com/T-B-F/pyHCA
https://doi.org/10.1101/249995 (preprint on bioRxiv)

Summary

Dark proteome (unannotated protein sequences) are made of foldable segments not related to domain families with some appearing to be disorder.

Deteced novel domains without homology signal that have a higher content of predicted disorder than ancient domains → “growth slow and moult model”

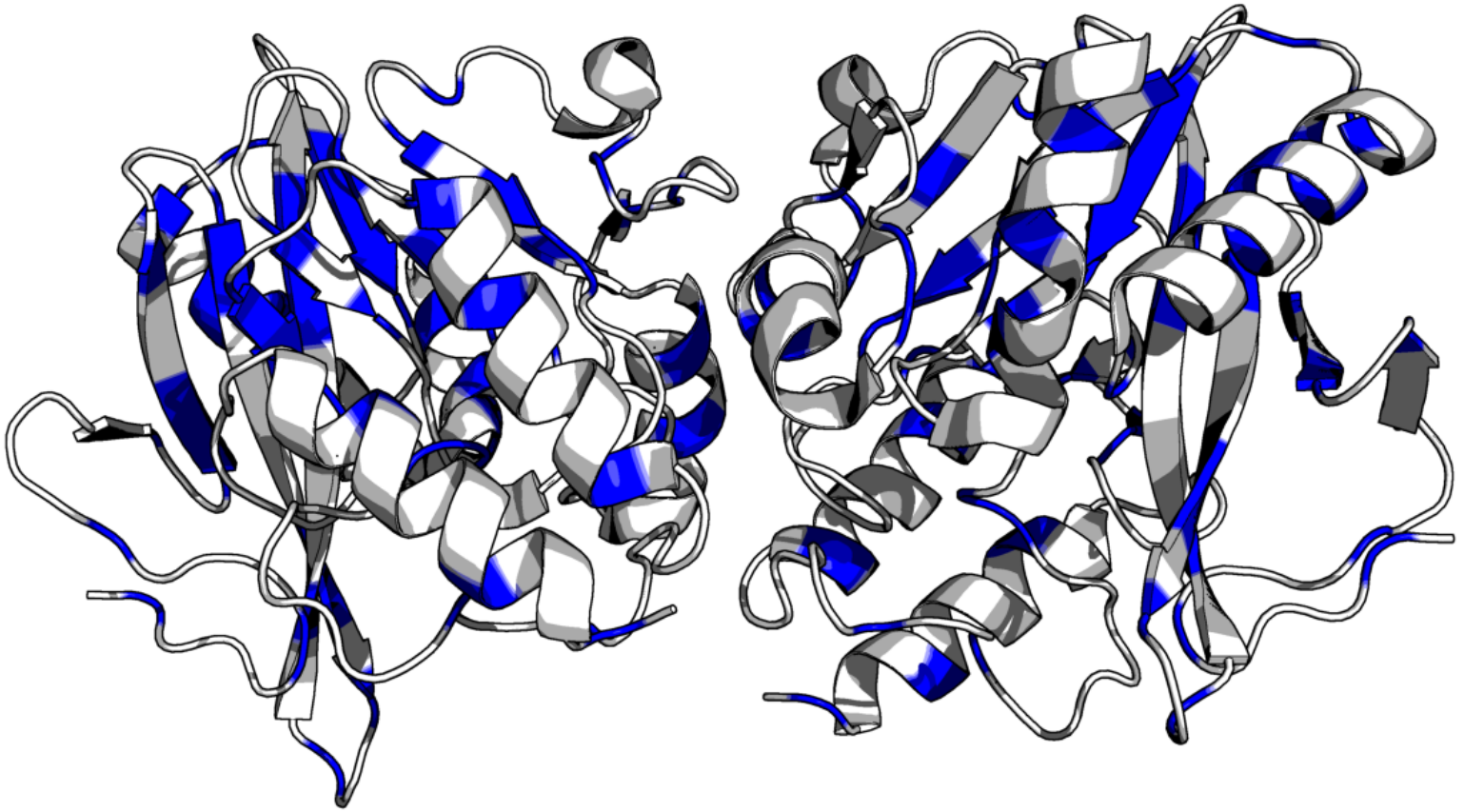
Thank you

IMPMC (Paris)
Isabelle Callebaut

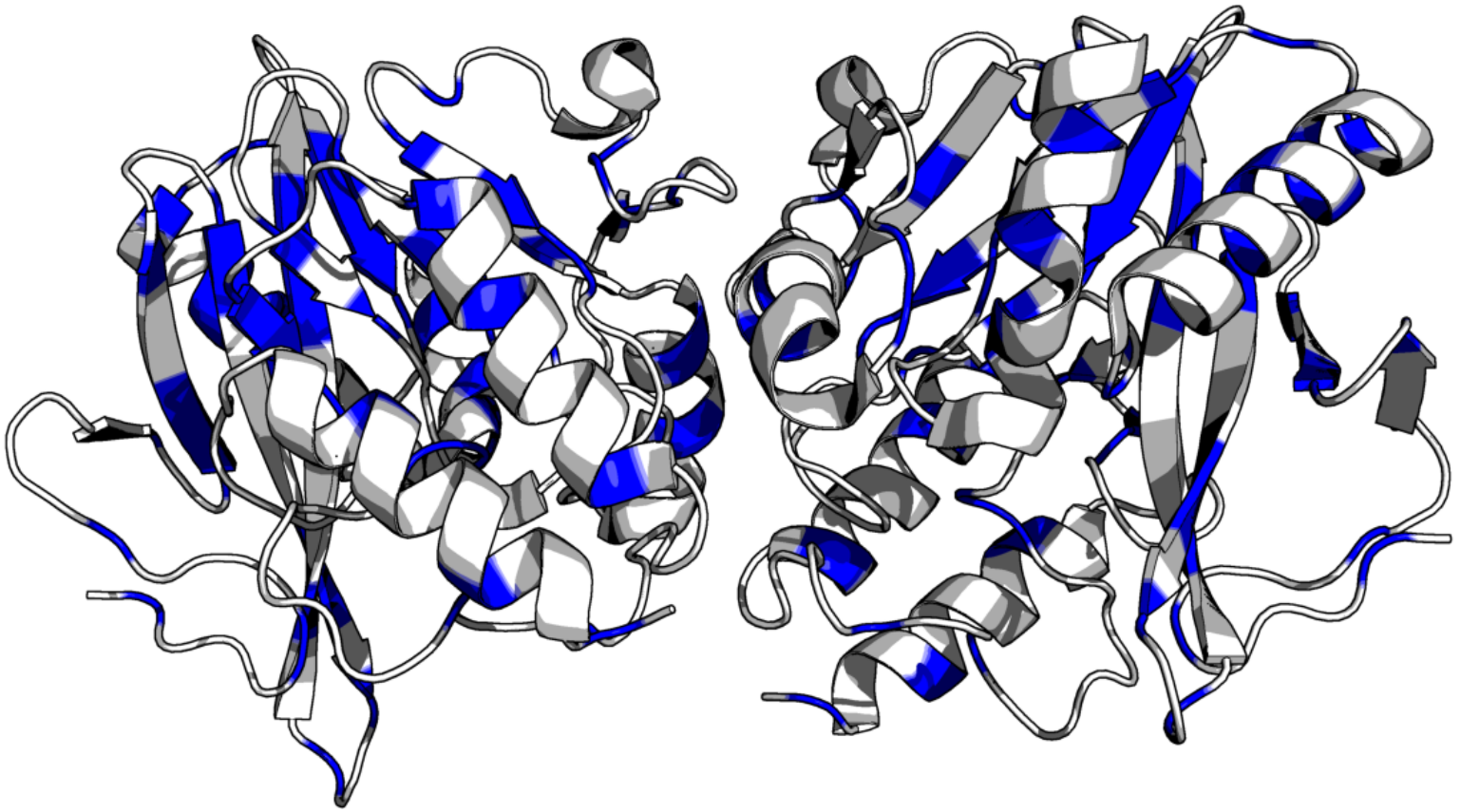
Evolutionary bioinformatics (Münster)
Erich Bornberg-Bauer
Steffen Klasberg

LCQB (Paris)
Alessandra Carbone

Hydrophobic Cluster Analysis

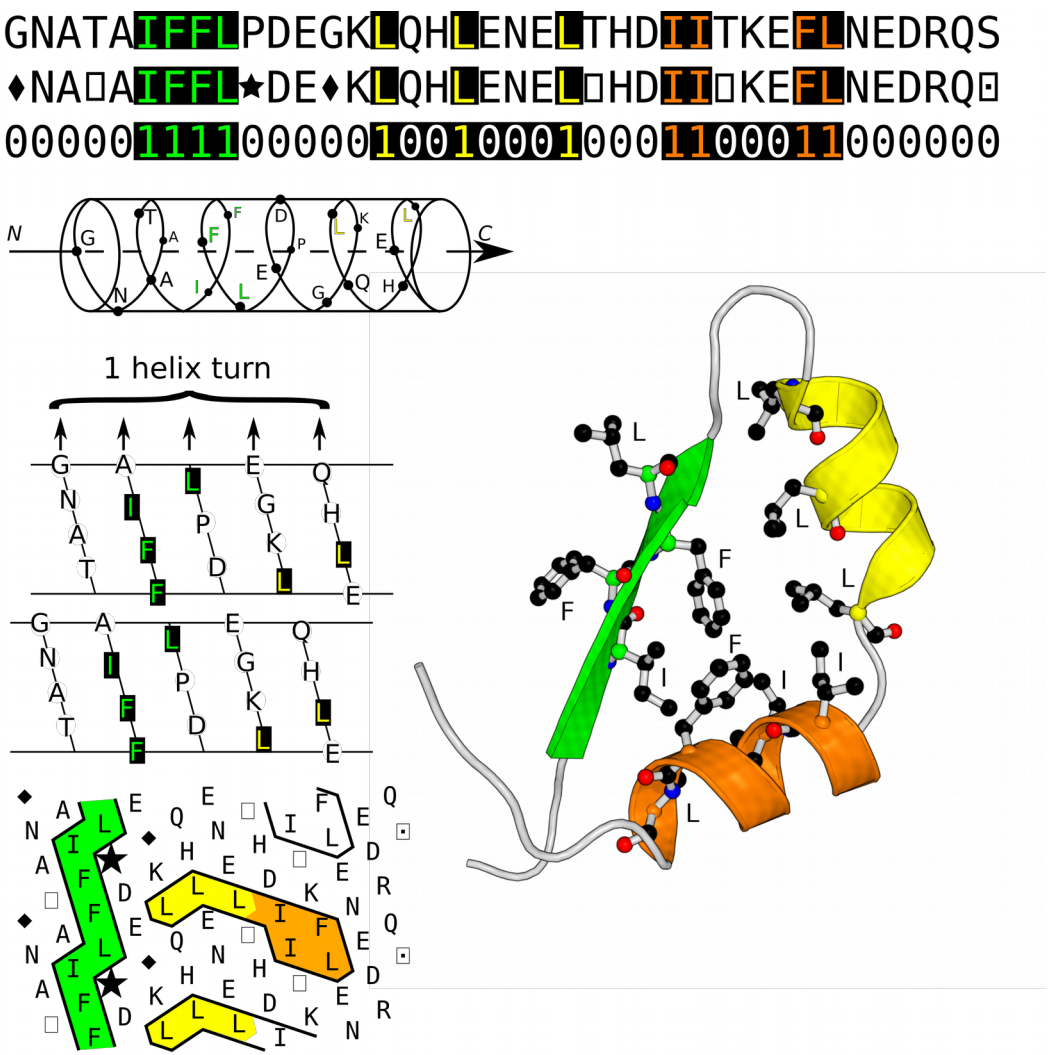


Hydrophobic Cluster Analysis



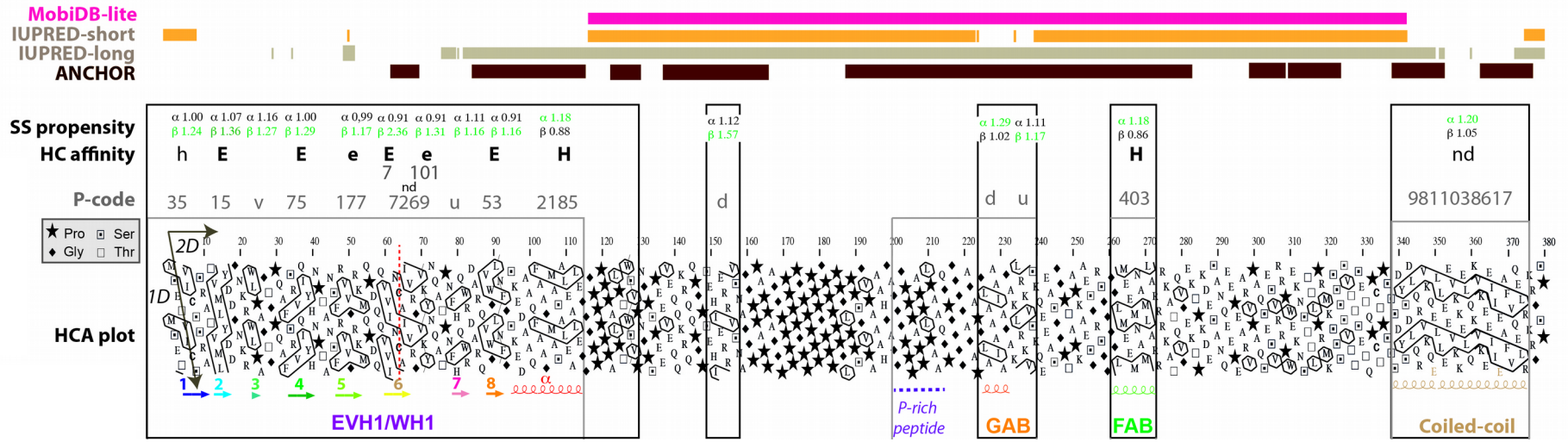
SNAIRPTIGQQMETGDQRFGDLVFRQLAPNVWQHTSYLDVPGFGAVASNGLIVRDGGRVLVVDTAWTDDQTAQILNWIKQ
000100010001000000100111001000110000110100100100001110000011110001000000001101100
EINLPVALAVVTHAHQDKMGGMDALHAAGIATYANALSNQLAPQEGMVAAQHSLTFAANGWVEPATAPNFGPLKVFYPGP
010101010110000000100100100001001000100010000011000001010000110000000010010111000
GHTSDNITVGIDGTDIAFGGCLIKDSKAKSLGNLGDADTEHYAASARAFGAAFPKASMIVMSHSAPDSRAAITHTARMAD
000000101010000101000110000000010010000000100000010001000011110000000000100000100
KLR
010

Hydrophobic Cluster Analysis



Hydrophobic Cluster Analysis

Enabled/vasodilator-stimulated phosphoprotein (Ena/VASP)



T. Bitard-Feildel et al – Proteomics *under review*

credits I. Callebaut