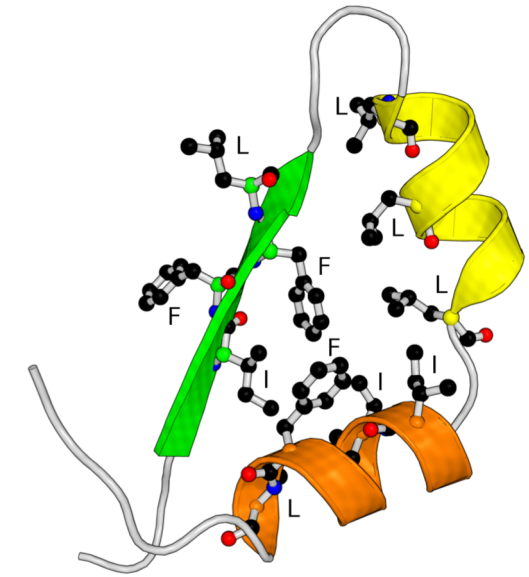
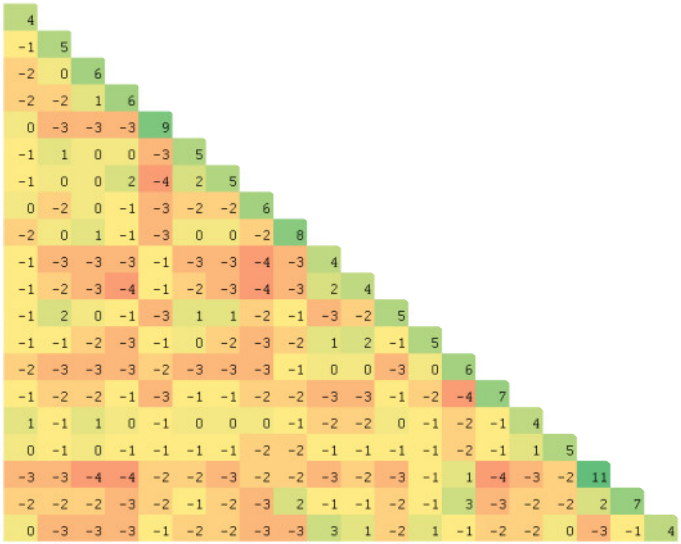
Evolution of structural signatures in families of soluble protein domains

Substitution matrix of hydrophobic clusters from sequence alignments

M2 BIM-BMC 2021-2022
UE GENOM

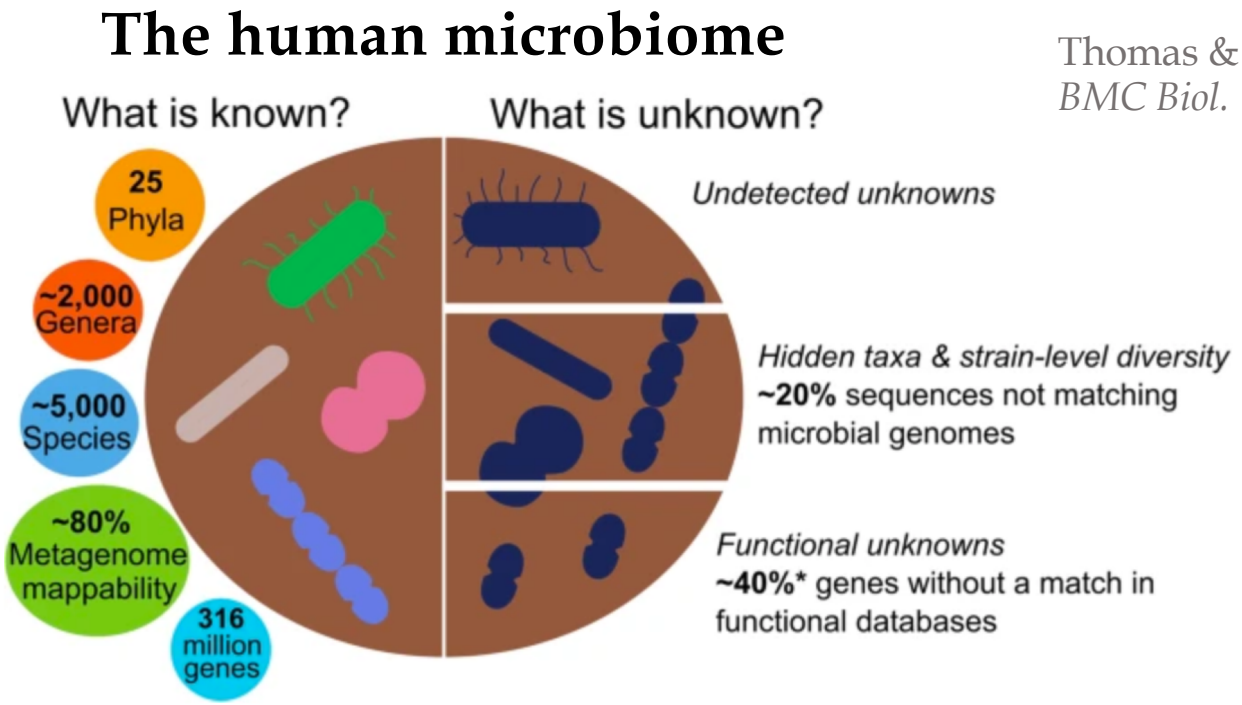




elodie.duprat@sorbonne-universite.fr

A lot of biological unknown

- are challenging the genome annotation scheme based on sequence similarity
- prevent to decipher the functional potential of genomes

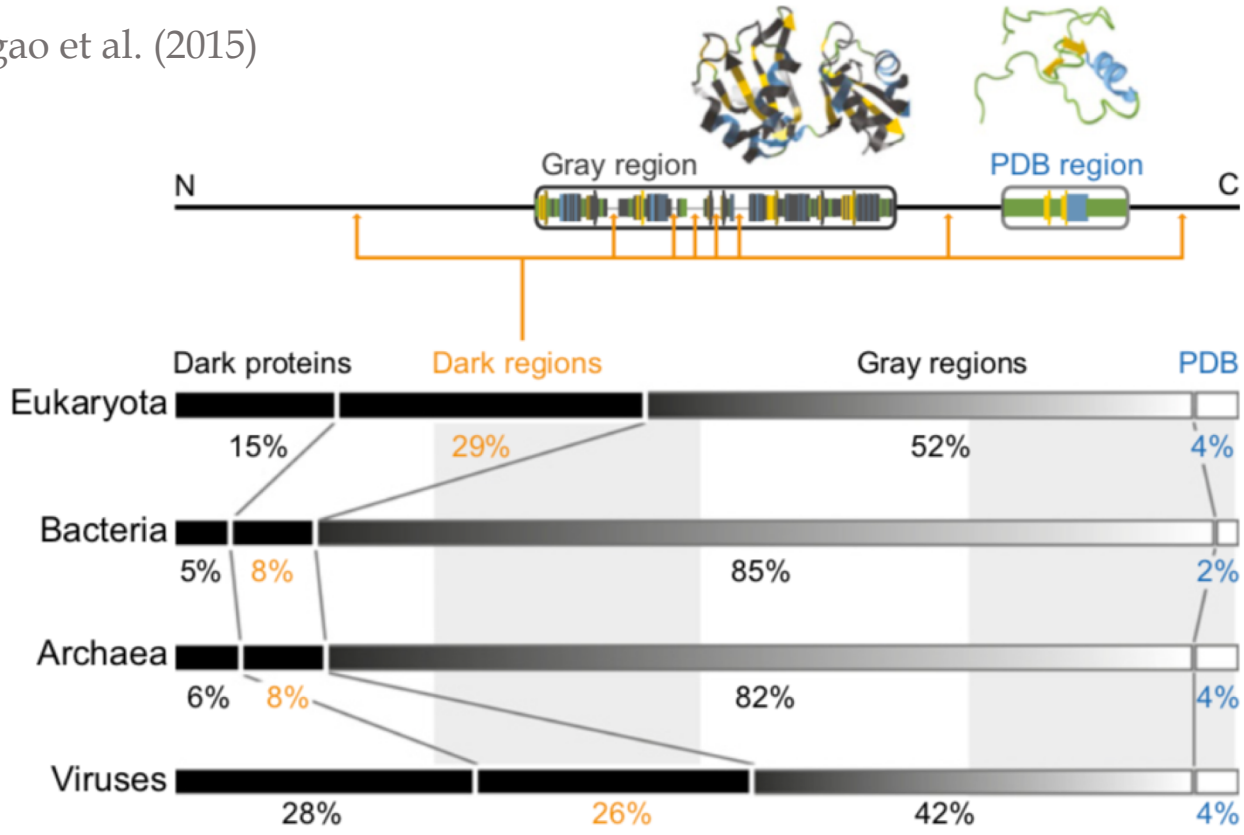


Thomas & Segata (2019)
BMC Biol.

A lot of biological unknown

- are challenging the genome annotation scheme based on sequence similarity
- prevent to decipher the functional potential of genomes

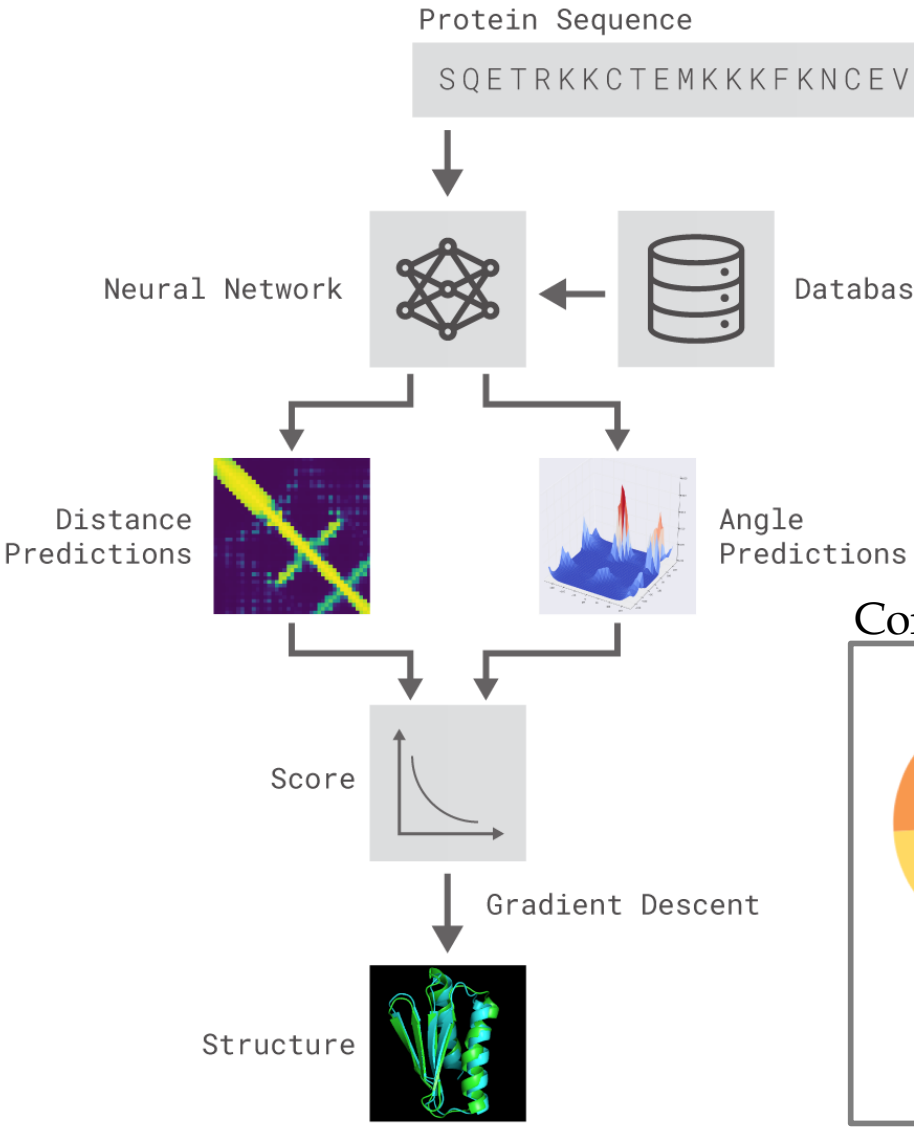
Perdigao et al. (2015)
PNAS



Swissprot
546 000 sequences

A lot of biological unknown

- are challenging the genome annotation scheme based on sequence similarity
- prevent to decipher the functional potential of genomes



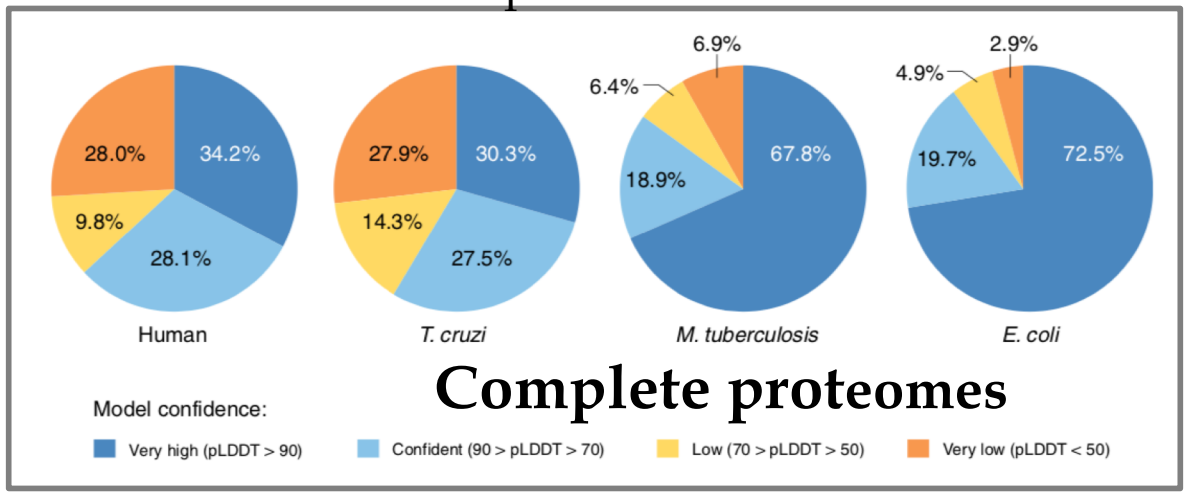
AlphaFold

Jumper et al. (2021) *Nature*
Tunyasuvunakool et al. (2021) *Nature*

Porta-Pardo et al. (2021) *bioRχiv*

Thornton et al. (2021) *Nat. Med.*

Confidence scores for AlphaFold models



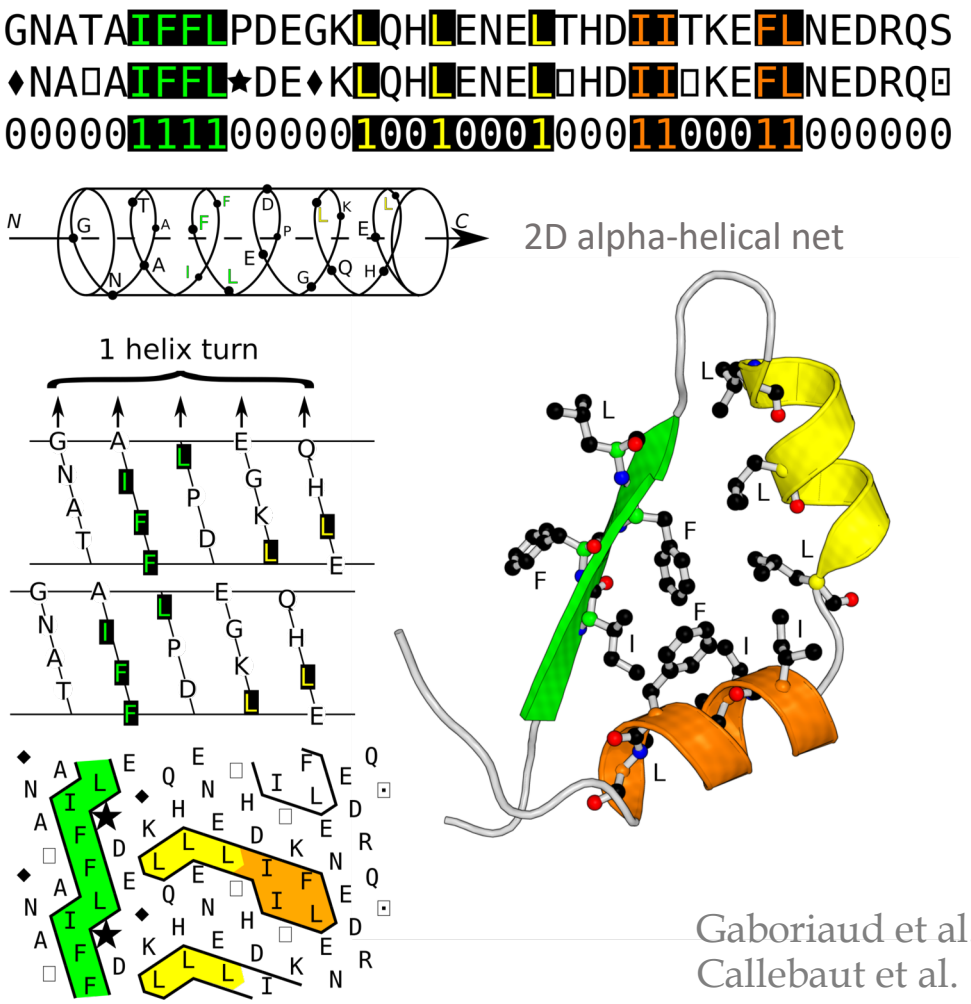
Complete proteomes

Structural signatures of soluble protein domains

- invariants of folds, more conserved than sequences
- Hydrophobic Clusters (HC):

(i) defined from amino acid sequence

(ii) match regular secondary structures



Gaboriaud et al. (1987) *FEBS Lett.*
Callebaut et al. (1997) *Cell Mol. Life Sci*

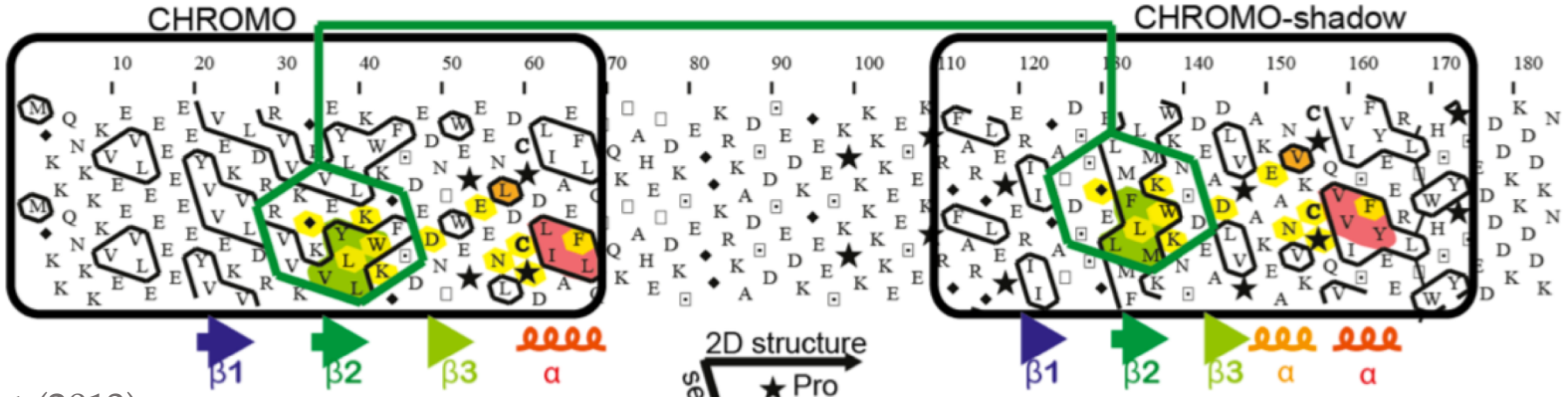
Structural signatures of soluble protein domains

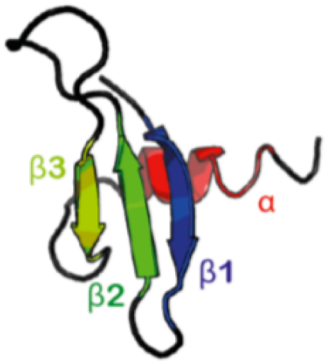
- invariants of folds, more conserved than sequences
- Hydrophobic Clusters (HC):

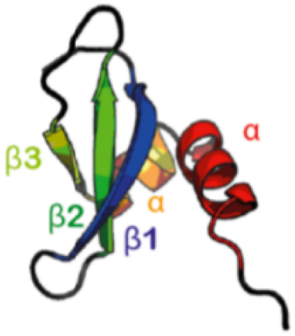
(i) defined from amino acid sequence

(ii) match regular secondary structures

→ Useful for the detection of remote sequence homology, foldable regions (even induced order in IDPs) and fold core







CBX1_MOUSE
UniProt: P83917
PDB: 1AP0, 1DZ1
19% seq. identity (N.S.)

Faure & Callebaut (2013)
Bioinformatics

Sangaré et al. (2016)
Nat. Comm.

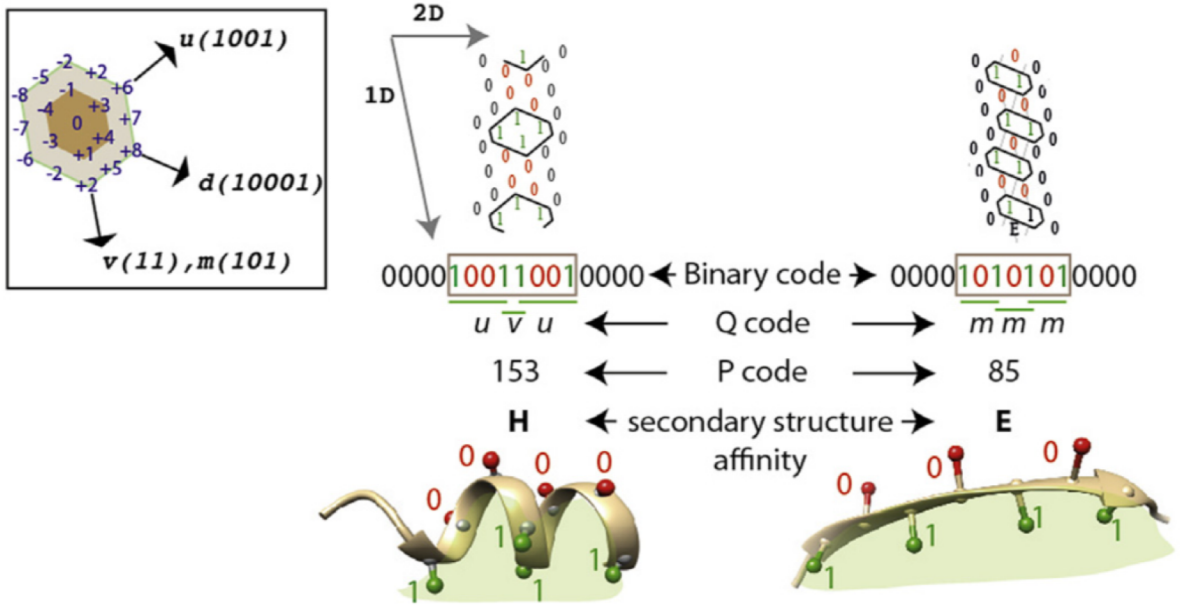
Faure et al. (2019)
Genome Biol. Evol.

Structural signatures of soluble protein domains

- invariants of folds, more conserved than sequences
- Hydrophobic Clusters (HC):
(i) defined from amino acid sequence
(ii) match regular secondary structures

→ Useful for the detection of remote sequence homology, foldable regions (even induced order in IDPs) and fold core

→ **Available database of signatures for 3D soluble domains (HCDB)**



Lamiable et al. (2019)
Biochimie

Structural signatures of soluble protein domains

- invariants of folds, more conserved than sequences
- Hydrophobic Clusters (HC):

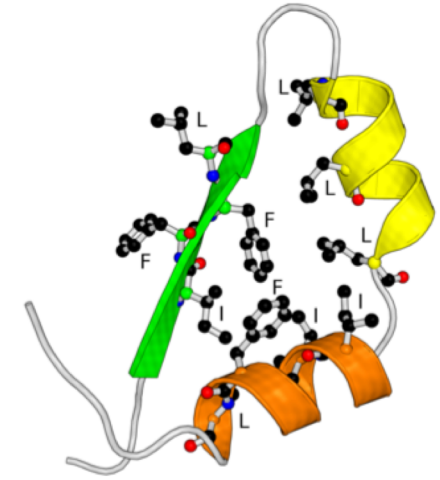
(i) defined from amino acid sequence

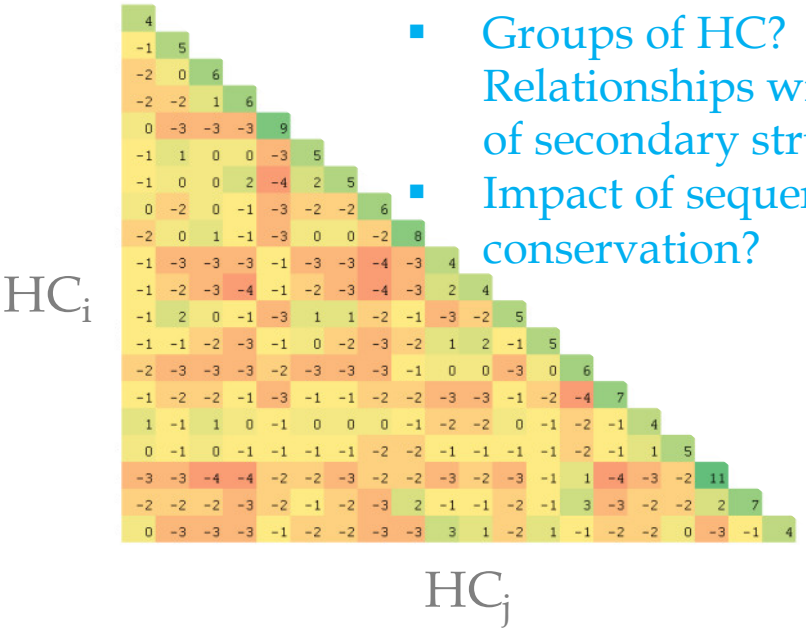
(ii) match regular secondary structures

→ Useful for the detection of remote sequence homology, foldable regions (even induced order in IDPs) and fold core

→ Available database of signatures for 3D soluble domains (HCDB)

→ **PROJECT:** *Build a HC substitution matrix from sequence alignments*





- Groups of HC? Relationships with types of secondary structures?
- Impact of sequence conservation?

Main steps of the project

Datasets

SOLUBLE DOMAINS (3D)

SCOPe

Release 2.08
– September 2021

Mapping file
(Pfam ftp)

FAMILIES OF
DOMAIN SEQUENCES
& ALIGNMENTS

Pfam

Pfam-A release 35.0
– November 2021

Multiple sequence alignments of families of soluble domains
(representative proteomes, e.g. RP15, RP35)

`https://scop.berkeley.edu/downloads/parse/dir.des.scope.2.08-stable.txt`

`https://ftp.ebi.ac.uk/pub/databases/Pfam/mappings/pdb_pfam_mapping.txt`

`https://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.rp??.gz`

Main steps of the project

Datasets

SOLUBLE DOMAINS (3D)

SCOPe

Release 2.08
– September 2021

Mapping file
(Pfam ftp)

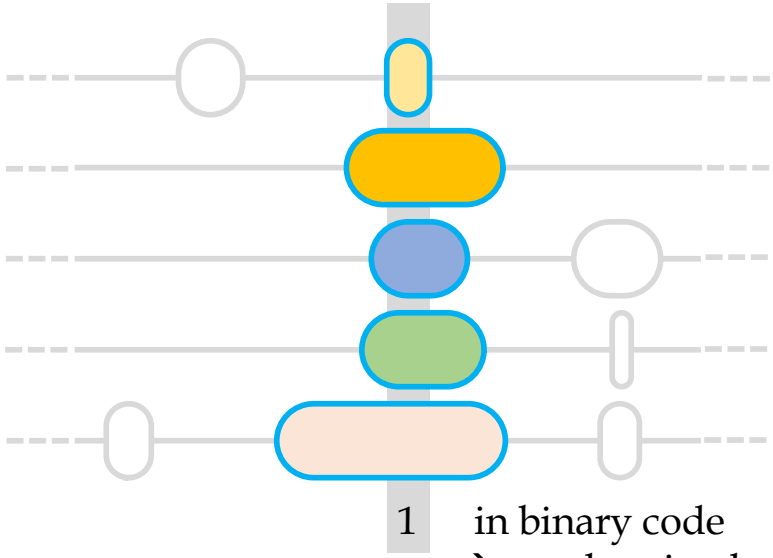
FAMILIES OF
DOMAIN SEQUENCES
& ALIGNMENTS

Pfam

Pfam-A release 35.0
– November 2021

Multiple sequence alignments of families of soluble domains
(representative proteomes, e.g. RP15, RP35)

Schematic representation of a set of homologous HC in a given family of soluble domains:



1 in binary code
→ overlapping hydrophobic clusters

- Binary encoding of aligned sequences
- Delimitation and encoding of HC – see Lamiable et al. 2019
- Identification of sets of homologous HC
- (Graph representation – HC as nodes)
- Building of a HC substitution matrix

HCDB → HC counts in soluble domains (Lamiable et al. 2019)