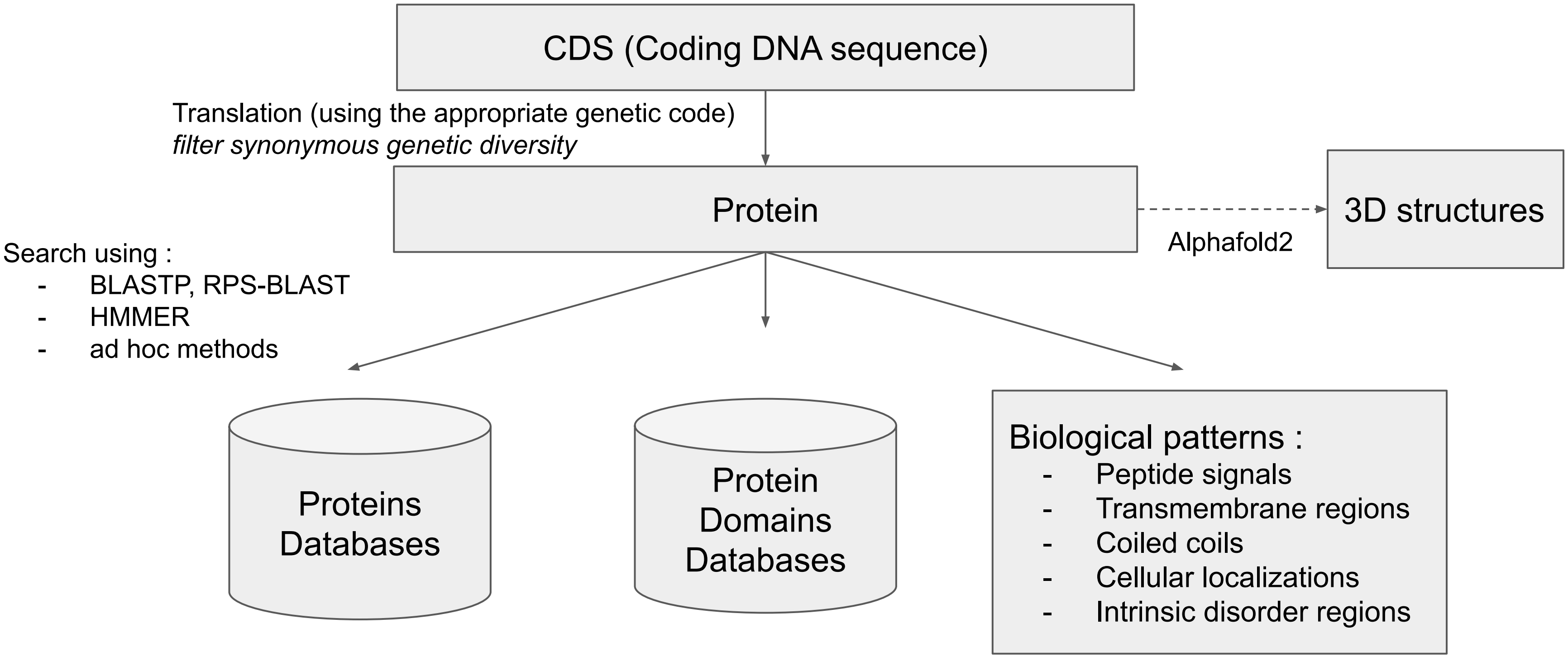
5. Introduction to functional annotation

Guillaume GAUTREAU, 28/09/2022

What is the functional role of an identified coding genes ?

Simplify analysis pipeline :



CDS (Coding DNA sequence)

Translation (using the appropriate genetic code)
filter synonymous genetic diversity

Protein

Alphafold2

3D structures

Search using :
- BLASTP, RPS-BLAST
- HMMER
- ad hoc methods

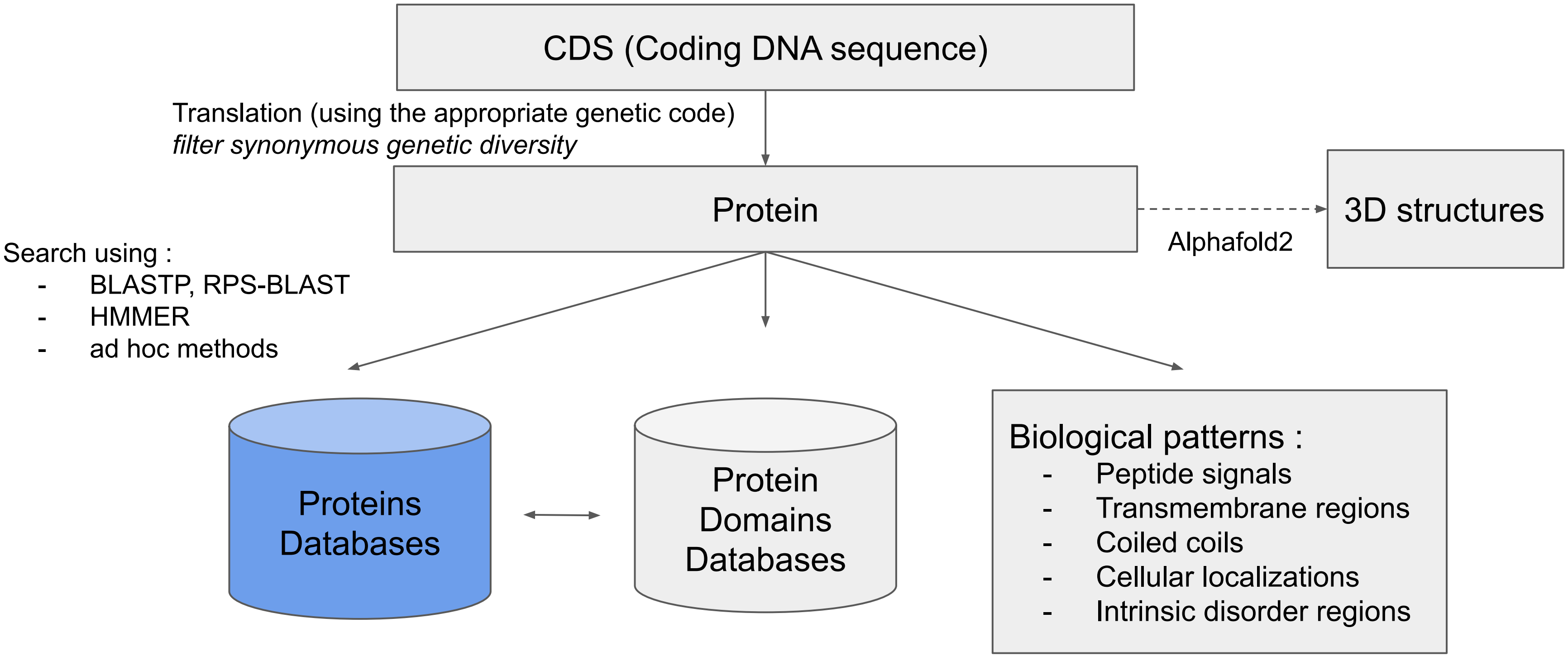
Proteins Databases

Protein Domains Databases

Biological patterns :
- Peptide signals
- Transmembrane regions
- Coiled coils
- Cellular localizations
- Intrinsic disorder regions

What is the functional role of an identified coding genes ?

Simplify analysis pipeline :



CDS (Coding DNA sequence)

Translation (using the appropriate genetic code)
filter synonymous genetic diversity

Protein

Alphafold2

3D structures

Search using :

- BLASTP, RPS-BLAST
- HMMER
- ad hoc methods

Proteins Databases

Protein Domains Databases

Biological patterns :

- Peptide signals
- Transmembrane regions
- Coiled coils
- Cellular localizations
- Intrinsic disorder regions

Protein databases

Each entry contained in those databases can correspond either :

- A specific protein
- A protein family corresponding to a set a similar* protein, generally summarized by a :
    - A consensus sequence or representative sequence
    - A protein HMM profile
    - A fingerprint (rules about the succession of domains)

*Similarity means homologous genes, which can be either (depending on databases):

- Paralogous genes
- Orthologous genes

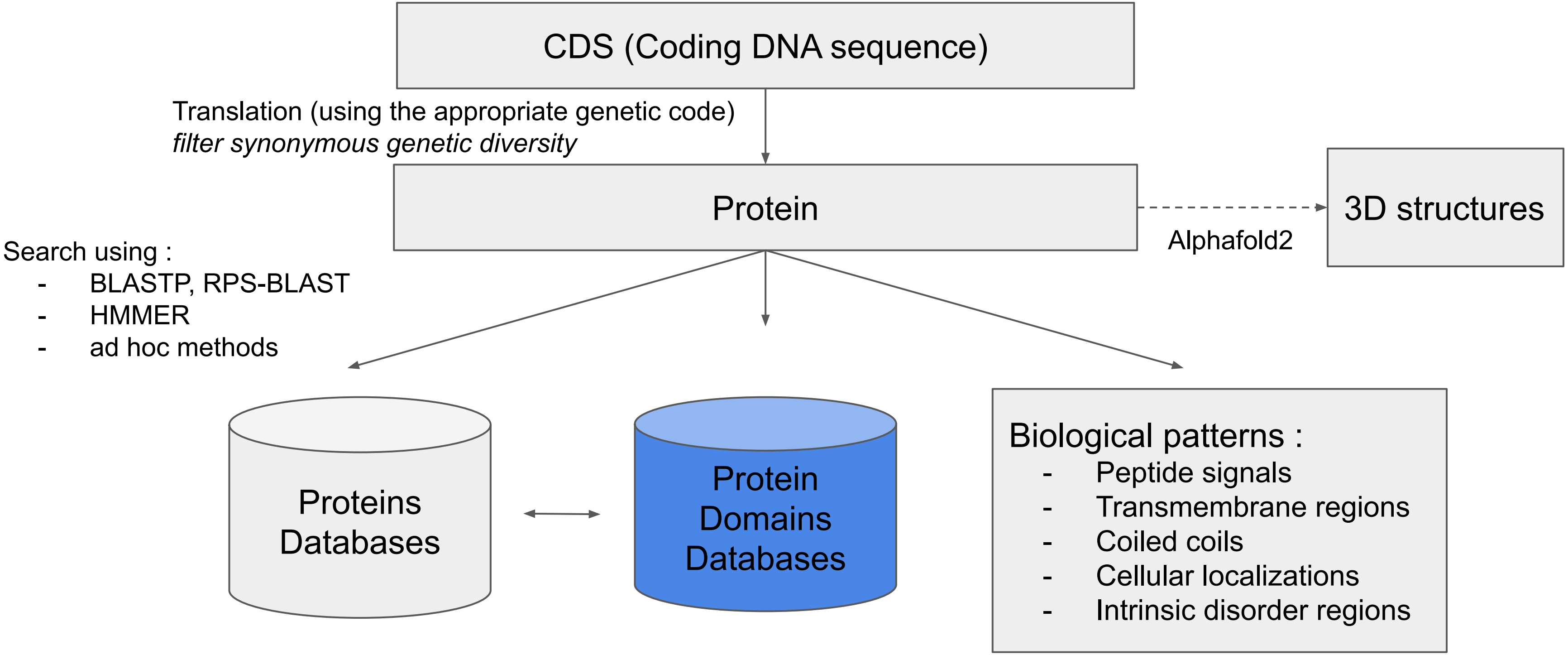
Protein databases

Some non-exhaustive examples of protein databases :

- Uniprot: protein sequences
    - TrEMBL: the translations of all coding sequences in GenBank (~200M)
    - Swissprot: manually curated part of TrEMBL (1%)
- TIGRFAMs: homologous proteins having a similar function (HMM profiles)
- HAMAP: set of manually curated protein
- PANTHER: homologous proteins having a similar function (HMM profiles)
- CATH-Gene3D: Describes protein families and domain architectures
- SUPERFAMILY: represent all proteins of a known structure (HMM profiles)
- SFLD: classification of enzymes (sequence-structure associated to chemical function)
- COG (orthologs): HMM profile
- EggNOG (orthologs): HMM profile or representative sequence
- Pfam: HMM profile of protein families
- KEGG orthologs

What is the functional role of an identified coding genes ?

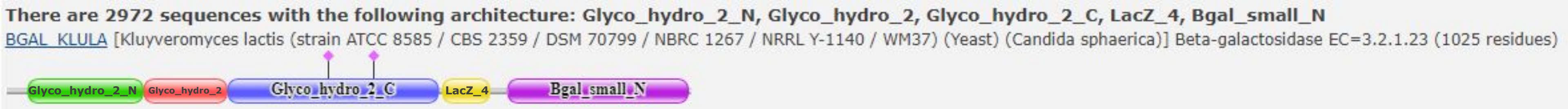
Simplify analysis pipeline :



Protein Domain Databases

Each entry contained in those databases correspond to a protein domain and their functional annotation :

example LacZ (5 domains):



There are 2972 sequences with the following architecture: Glyco_hydro_2_N, Glyco_hydro_2, Glyco_hydro_2_C, LacZ_4, Bgal_small_N

BGAL_KLULA [Kluyveromyces lactis (strain ATCC 8585 / CBS 2359 / DSM 70799 / NBRC 1267 / NRRL Y-1140 / WM37) (Yeast) (Candida sphaerica)] Beta-galactosidase EC=3.2.1.23 (1025 residues)

Reminder: a protein domain is a region of a protein's polypeptide chain that is self-stabilizing and that folds independently from the rest of the protein.

Some domains can be detected but without any associated function :

- called DUF: Domain of Unknown Function

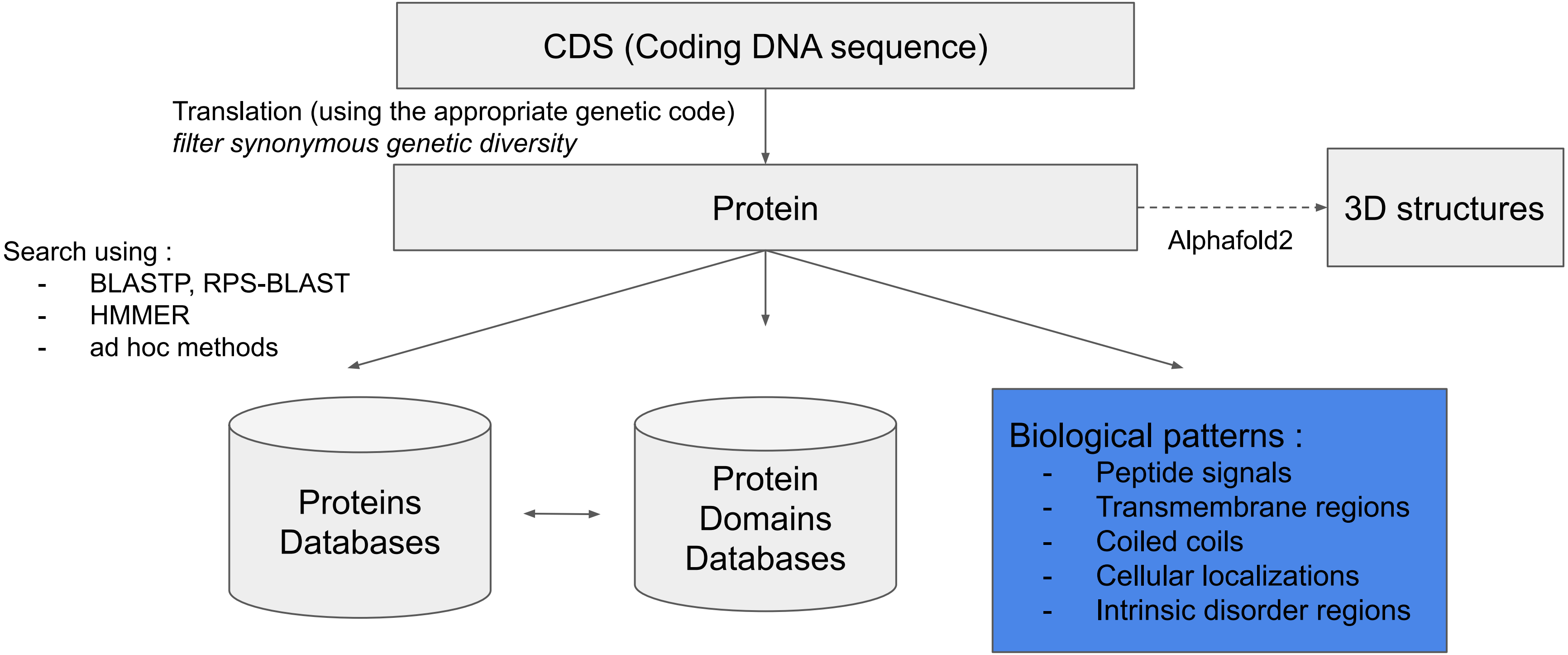
Protein Domain Databases

Some non-exhaustive examples of protein domain databases :

- PROSITE: biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs.
- SMART: identification and annotation of genetically mobile domains and the analysis of domain architectures
- CDD: Conserved Domain Database is a protein annotation resource that consists of a collection of annotated multiple sequence alignment models for ancient domains and full-length proteins.
    - Stored as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST.
- Pfam domain: HMM profile of Pfam domain

What is the functional role of an identified coding genes ?

Simplify analysis pipeline :



CDS (Coding DNA sequence)

Translation (using the appropriate genetic code)
filter synonymous genetic diversity

Protein

Alphafold2

3D structures

Search using :

- BLASTP, RPS-BLAST
- HMMER
- ad hoc methods

Proteins Databases

Protein Domains Databases

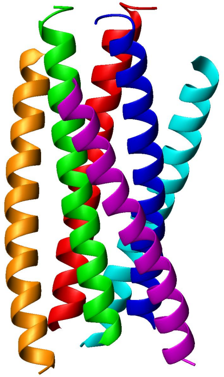
Biological patterns :

- Peptide signals
- Transmembrane regions
- Coiled coils
- Cellular localizations
- Intrinsic disorder regions

Biological patterns

Some patterns which can be recognized by dedicated methods :

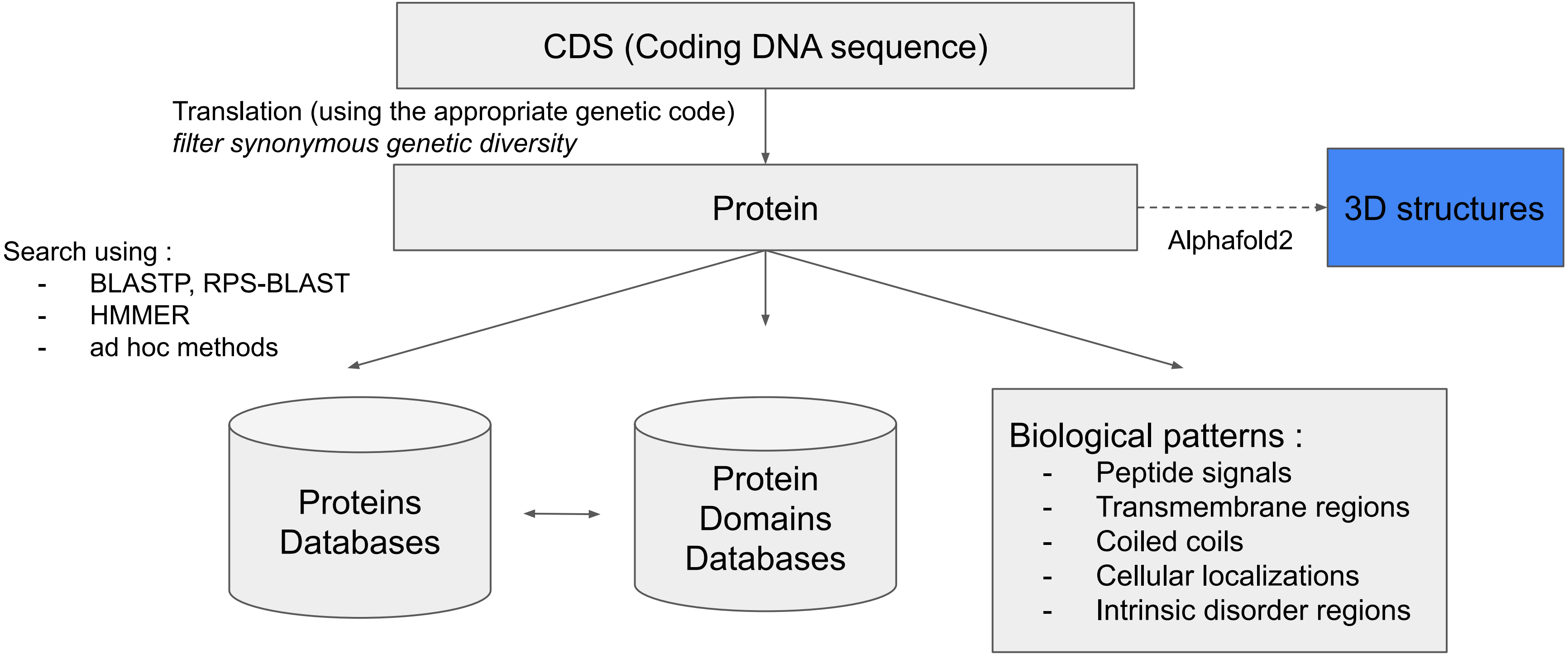
- Signal peptides and their cleavage sites : signalP (and Phobius)
- Transmembrane region: Phobius and TMHMM
- Coiled coils : Coils →
- Cellular localizations : PSORTb
    - Cytoplasmic
    - Cytoplasmic Membrane
    - Cell wall
    - Extracellular : flagellar, fimbrial, type III secretion apparatus, host-associated, spore
- Intrinsically disordered regions (less frequent in prokaryotes):
    - region structurally instable, and so very plastic, helping to:
        - protein interaction
        - protein scaffolding
        - post-transcriptional modification
    - Can be predicted using the MobiDB database



credit: Wikipedia

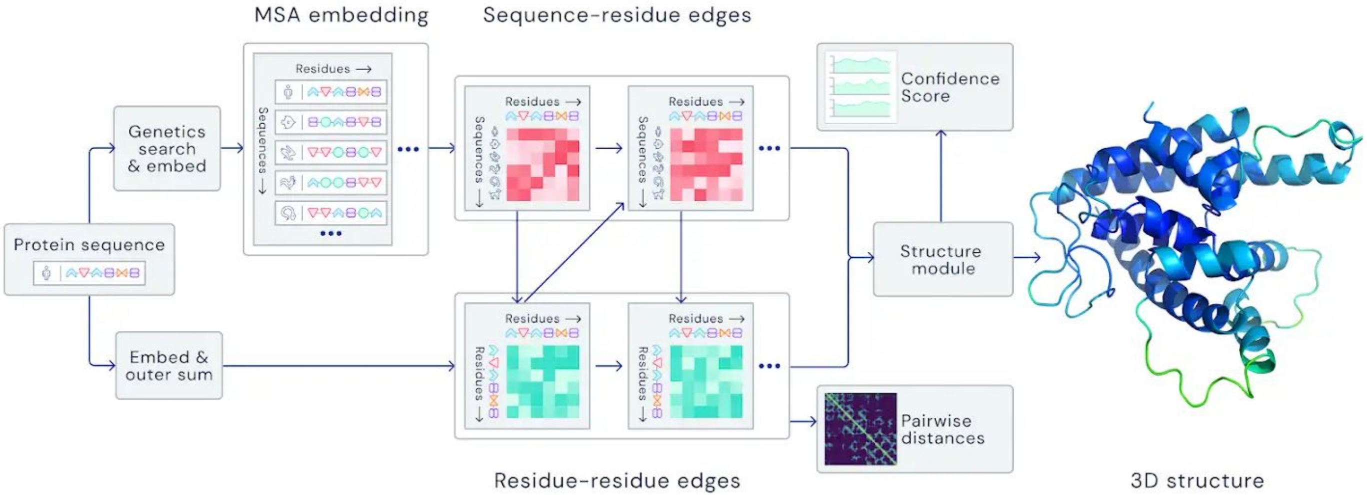
What is the functional role of an identified coding genes ?

Simplify analysis pipeline :



3D protein structures

- Protein Data Bank (PDB)
- AlphaFold2: a recent revolution in *ab initio* structural prediction



credit: https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

Conclusion

- Many of the presented tools are wrapped in InterProScan and EggNOG (lesson 8. on EggNOG et InterProScan)
- Almost all of presented databases are interconnected : protein, domain, biological pattern, 3D structures
- Many limits about these databases:
  - Many redundant informations
  - Often contradictory annotations or incomplete annotations (promiscuity activity, many active sites)
  - ⇒ many levels of precision/confidence with :
    - curated information (low number)
      versus
    - automatic annotations
  - Still a lot of unknown proteins or DUF
  - Computationally intensive
  - Still isolated annotations, must be improved by relational annotation
- Some approaches try to integrate all of these annotations using human reviewed rules to improve the quality of annotation (example : Unifire/Unirule based on InterProScan Results)