ETH zürich

551-1119-00L Microbial Community Genomics

Lecture:

Methods on computational genome mining

GOAL

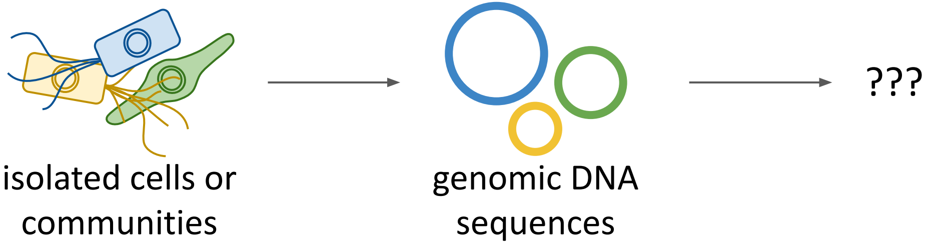
Understand the **different sequence algorithms** that can be applied to **genomes** providing **mechanistic** and **functional** information about the biological system

1. Introduction: the genomics rationale
2. Gene annotation
3. *Ab initio* gene annotation
    a. Sequence content
    b. Genetic elements
    c. Evaluation of sequence motifs
4. Evolutionary conservation: sequence alignment
5. Evolutionary conservation: HMMs
6. Sequence features for annotation
7. Structures biology
8. Closing remarks

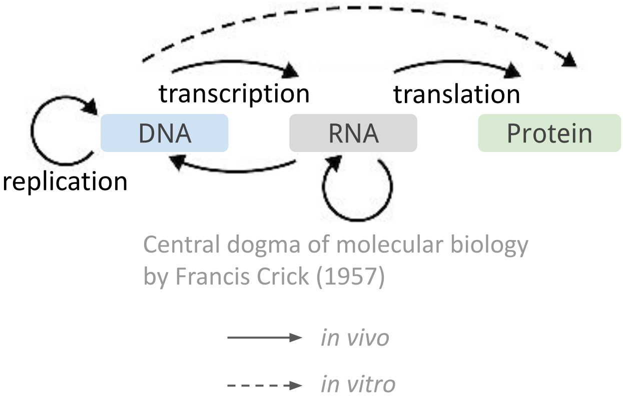
1. Introduction || The genomics rationale

1. Introduction || The genomics rationale

What sequencing provides so far:



Genomes are a valuable source of **biological** and **functional** information
The final goal in a genomics study usually covers the **genotype** ↔ **phenotype**



Central dogma of molecular biology
by Francis Crick (1957)

polymer	Deoxyribonucleic acids (DNA) - ACGT	Ribonucleic acids (RNA) - ACGU	20 amino acids
unit	gene	mRNA, tRNA, rRNA, ncRNA…	protein
set	genome	transcriptome	proteome
function	information	intermediary & regulation	structural & biochemical

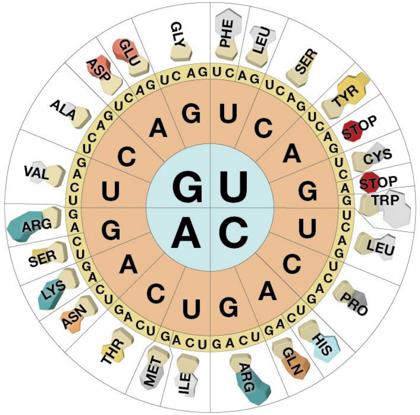
Understanding these processes allow to understand **regulation** and **function** in organisms (transcriptome and proteome) from **genomic** information

2. Gene annotation || ORF scanning

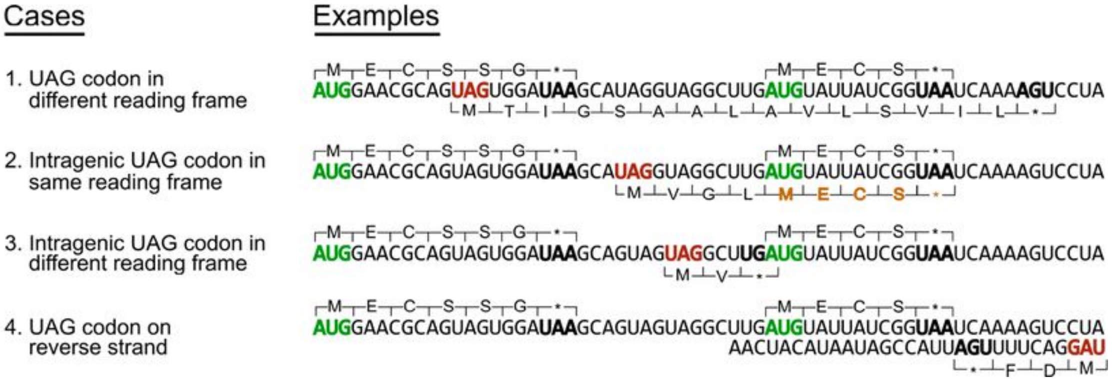
2. Gene annotation | | ORF scanning

Gene annotation is a primary step that relies on the “Open Reading Frame (**ORF**) **scanning**” process:

1. 6 ORFs in a genome → why this number?
    - To cover 20 amino acids + stop → 4^1; 4^2; $4^3 = 64$
        - Evolutionary process selecting the codons as nucleotide triplets
    - Genetic code = correspondence between codon - aa
        - It is “universal” and “degenerate”
2. Looking for every start-stop codon sequence:
    - Start codon encode for methionine (e.g. AUG)
    - Stop codon block translation (e.g. UAG, UAA, UGA)



Genetic code in a codon table



Cases | Examples

1. UAG codon in different reading frame

```
┌M┬E┬C┬S┬S┬G┬•┐                     ┌M┬E┬C┬S┬•┐
AUGGAACGCAGUAGUGGAUAAGCAUAGGUAGGCUUGAUGUAUUAUCGGUAAUCAAAAGUCCUA
          └M┴T┴I┴G┴S┴A┴A┴L┴A┴V┴L┴S┴V┴I┴L┴•┘
```

2. Intragenic UAG codon in same reading frame

```
┌M┬E┬C┬S┬S┬G┬•┐                     ┌M┬E┬C┬S┬•┐
AUGGAACGCAGUAGUGGAUAAGCAUAGGUAGGCUUGAUGUAUUAUCGGUAAUCAAAAGUCCUA
                        └M┴V┴G┴L┴M┴E┴C┴S┴•┘
```

3. Intragenic UAG codon in different reading frame

```
┌M┬E┬C┬S┬S┬G┬•┐                     ┌M┬E┬C┬S┬•┐
AUGGAACGCAGUAGUGGAUAAGCAGUAGUAGGCUUGAUGUAUUAUCGGUAAUCAAAAGUCCUA
                              └M┴V┴•┘
```

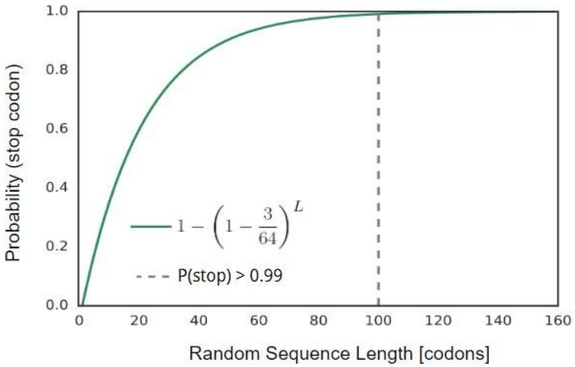
4. UAG codon on reverse strand

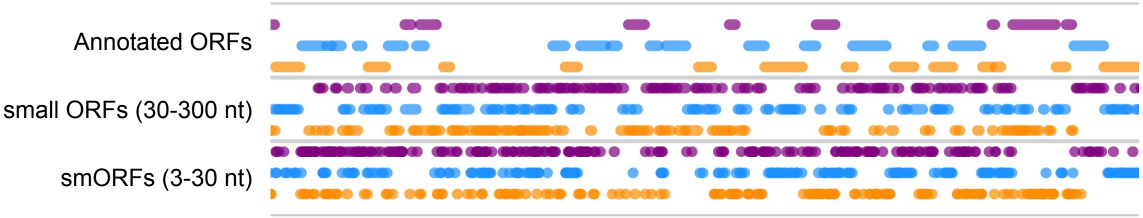
```
┌M┬E┬C┬S┬S┬G┬•┐                     ┌M┬E┬C┬S┬•┐
AUGGAACGCAGUAGUGGAUAAGCAGUAGUAGGCUUGAUGUAUUAUCGGUAAUCAAAAGUCCUA
                                  AACUACAUAAUAGCCAUUAGUUUUCAGGAU
                                                   └•┴F┴D┴M┘
```

Complex for humans, very easy task for a computer

2. Gene annotation | | ORF scanning

3. Any sequence larger than 300 nucleotides can be considered to be a gene



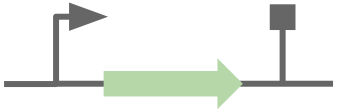


ORFs found in 10Kb (10,000 DNA bases) in a *Mycoplasma pneumoniae* bacterial species

4. Additional **features** need to be considered to accurately annotate every gene
   - Genes that are smaller than 300 nt are tricky, as there are many more ORFs than protein-coding ORF sequences (referred to as 'CDS'). For example, antimicrobial and signalling proteins tend to be ≤100 aa
   - Even large ORFs could not be encoding for proteins, for example:
     - long non-coding RNAs
     - pseudogenes → when a protein-coding gene regulation is mutated (no expression) or it translates to a non-functional protein due to mutations

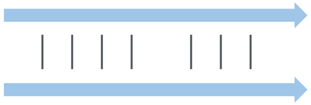
Which features can we consider?

2. Gene annotation | | Software tools approaches



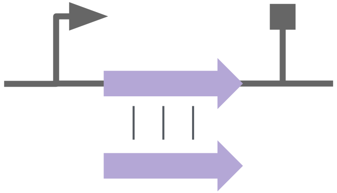
Ab initio:

- Sequence content (SC) comparative between **Coding** vs. **non-coding** in terms of:
  - GC content, Codon Adaptation…
- Genetic **signals**
  - Promoters, Ribosome Binding Sites…
  - Alternative splicing (only eukaryotes)



Sequence Homology (SH):

- Coding sequences are **conserved**
- **Alignment** against DBs (known genes, expressed RNAs, function clusters...)



Combination (CM):

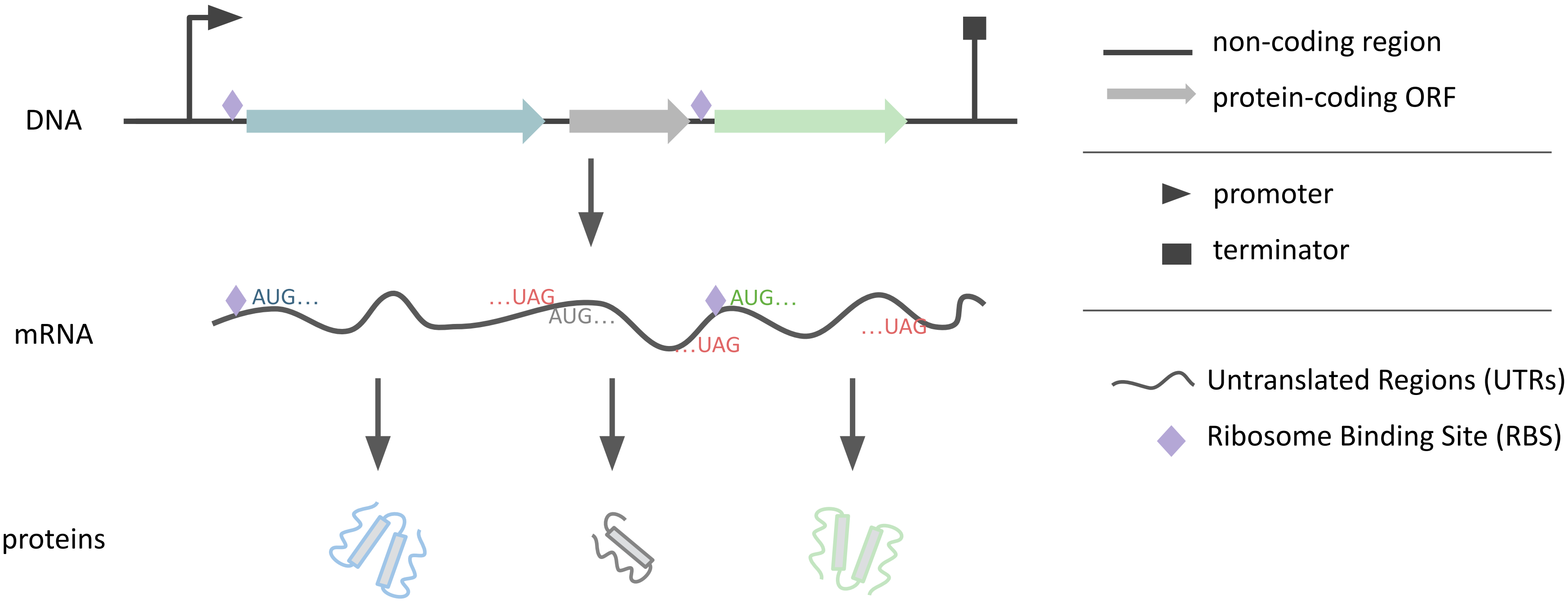
- Widely used in **genomic databases**
- NCBI Prokaryotic Genome Annotation Pipeline (PGAP) is the tool that runs when we submit a genome to NCBI

Tool	Year	Type	Signals	Dependencies
GeneMark	1992	SC	-	-
GeneMark.hmm	1998	SC	-	-
Glimmer	1998	SC	-	-
ORPHEUS	1998	CM	RBS	DPS alignments
BLAST	1999	SH	-	-
COGs	2001	SH	-	-
AMIGene	2003	SC	-	-
GeneMarkS	2005	SC	5'-UTR	-
BASys	2005	CM		Glimmer, BLAST
Glimmer3	2007	SC	RBS	-
ProtClustDB	2009	SH	-	BLAST
Prodigal	2010	SC	RBS	-
FGENESB	2011	SC	-	-
Prokka	2014	CM	RBS	Prodigal, BLAST
ZCURVE	2015	SC	RBS	-
PGAP	2016	CM	RBS	BLAST, COGs, ProtClustDB, Glimmer, GeneMarkS
CPC2	2017	CM	RBS	BLAST

3. *Ab initio* || Annotation “from the beginning”

3. *Ab initio* | | Annotation “from the beginning”

Genes are not expressed by default, they are often **regulated** by different sequence elements



Sequence content and genetic features can all be explored at the DNA level and provide additional genetic insights

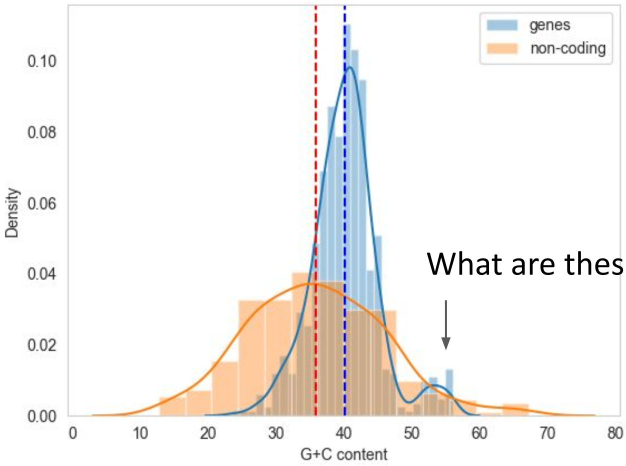
3. *Ab initio* | | Sequence composition: GC content

General idea: sequence composition differs between coding and non-coding regions

- Evolutionary biases can be used to distinguish genes in a genome
  - **Non-coding** regions will present **'random'** nucleotide compositions
  - **Coding** regions will **bias** towards combinations of **nucleotides** that give required amino acids in **proteins**

G + C content (also referred as GC%) describes the guanine and cytosine content of a biological sequence and has historically been reported to range **between 25% and 75%** for bacterial genomes

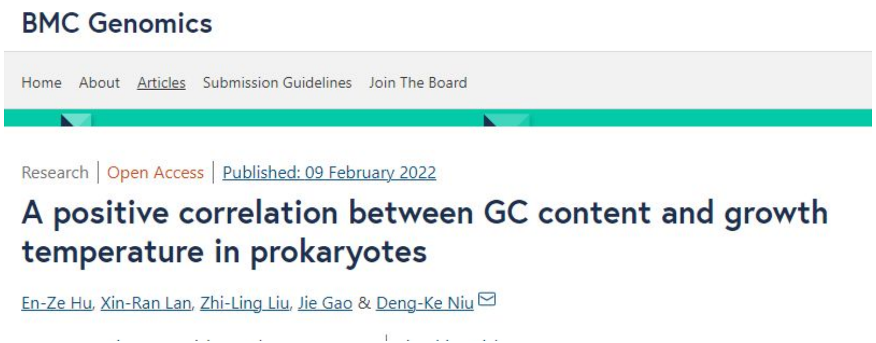
- GC% varies between coding and non-coding regions



What are these genes?

Example of *M. pneumoniae*, a low GC content organism

Other implications:



BMC Genomics

Research | Open Access | Published: 09 February 2022

A positive correlation between GC content and growth temperature in prokaryotes

En-Ze Hu, Xin-Ran Lan, Zhi-Ling Liu, Jie Gao & Deng-Ke Niu

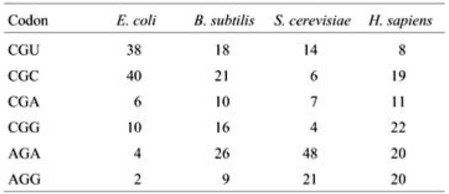
Not trivial to extrapolate mechanistic features…

3. *Ab initio* | | Sequence composition: Codon composition

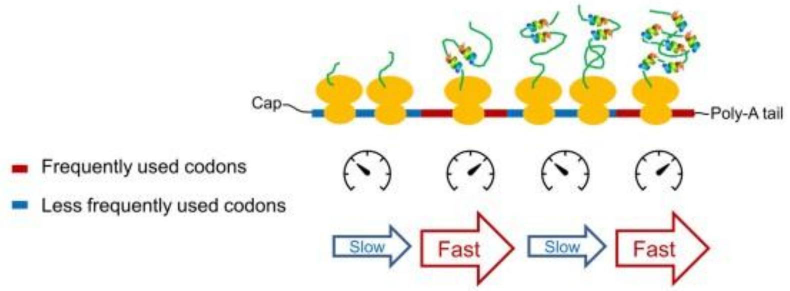
Codon usage bias refers to differences in the **frequency of occurrence of synonymous codons** in coding DNA

- Codon Adaptation Index (**CAI**) is a metric for codon biases that uses a set of reference genes in an organism, generally highly expressed, to measure how well other genes follow the same trend
  - Non-coding regions will present low CAIs
- Strong **correlation** with **GC%** and **tRNAs** abundances
- **Mechanistic** implications

Arg codon frequencies in 4 model organisms

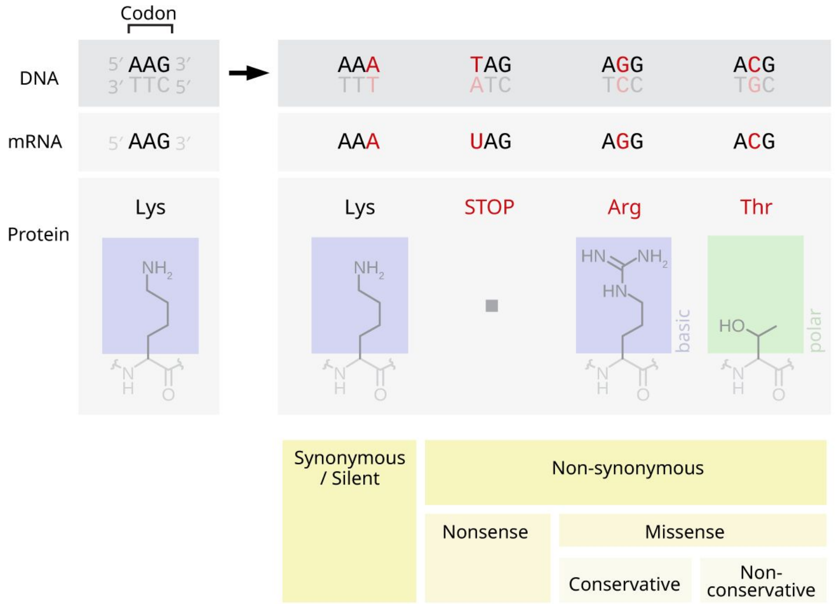


Codon	*E. coli*	*B. subtilis*	*S. cerevisiae*	*H. sapiens*
CGU	38	18	14	8
CGC	40	21	6	19
CGA	6	10	7	11
CGG	10	16	4	22
AGA	4	26	48	20
AGG	2	9	21	20



From: Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding - Molecular Cell (2015)

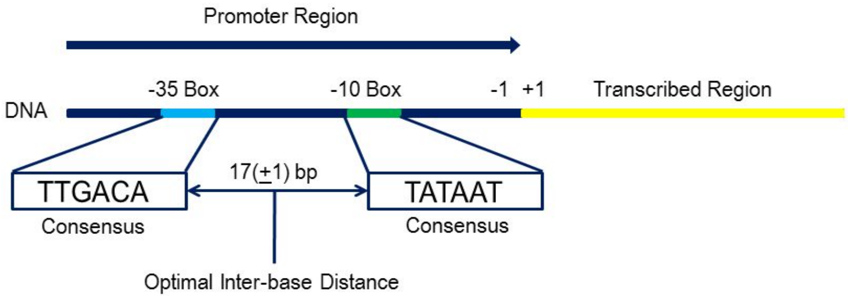
Mutation types:

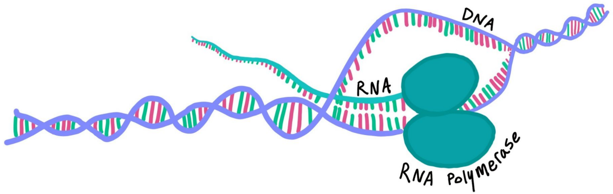


Note: mutations in the 3rd base of codon tend to be less ‘harmful’ as rarely induce nonsense mutation. The opposite happens with the 1st base.

3. *Ab initio* || Regulatory elements example: Promoters

A **promoter** is a sequence of DNA to which proteins bind to initiate transcription of a single RNA transcript

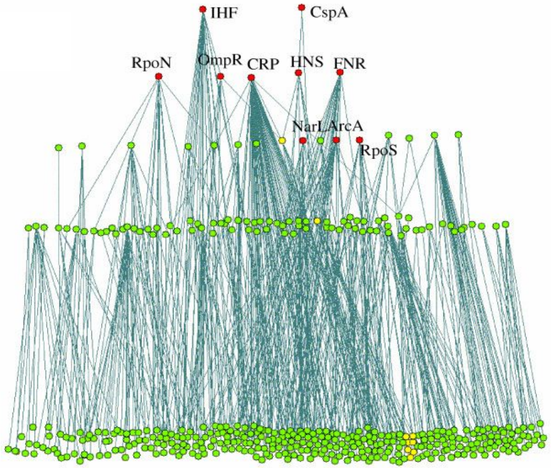




Promoters regulate downstream ≥1 protein-coding genes and also functional RNAs

- Genes expressed under the same promoter → **operon**
  - Corregulation of similar functions
    - 1st example Lac operon by Jacob & Monod

Additionally, there are several **Transcription Factors (TFs)** that can modulate the coexpression of different genes even if they are not in the same operon



Transcriptional regulatory network in *Escherichia coli*

Certain TF are active under specific conditions (e.g., cold-shock, heat-shock, osmotic stress…)

SAPPHIRE (kuleuven.be)
BPROM - Prediction of bacterial promoters (softberry.com)
Online Analysis Tools - Promoters (molbiol-tools.ca)

3. *Ab initio* || Evaluating sequence motifs

A **Position Weight Matrix** (PWM) quantitatively evaluates how well a given sequence matches a given sequence “motif”. These can include:

- Promoters: TATAAT (also referred to as TATA-box or Pribnow sequence)
  - Each transcription factor have a specific sequence motif as well
- Terminators:
  - Hairpin (measured by RNA folding) + poly-U
  - Rho binding sites
- RBS: AGGAGG (Shine-Dalgarno motif)

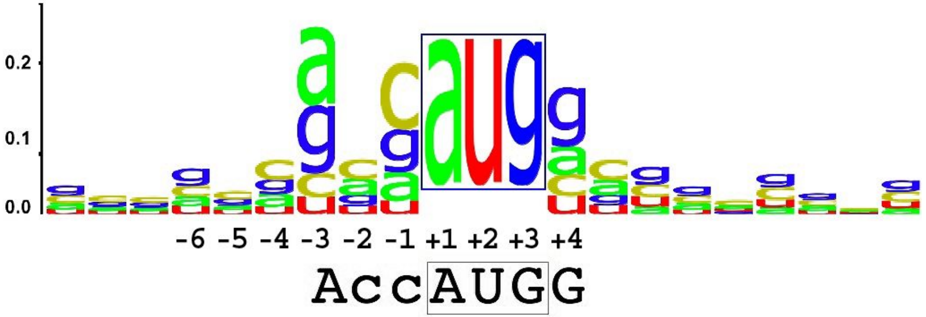
- These motifs may vary between species → evolution as driving force
- **Distance** between the regulatory motif and the regulated gene also matters

A	0.1	0.8	0	0.7	0.5	0
C	0	0.1	0.3	0.1	0.2	0.3
G	0	0	0.2	0.1	0.1	0.1
T	0.9	0.1	0.5	0.1	0	0.6

Product accumulated score:

```
TATAAT = 0.076
TACCCT = 0.002
CAACTT = 0
```

Bit score logos can be used to graphically represent a motif



This same approach works with amino acid sequences

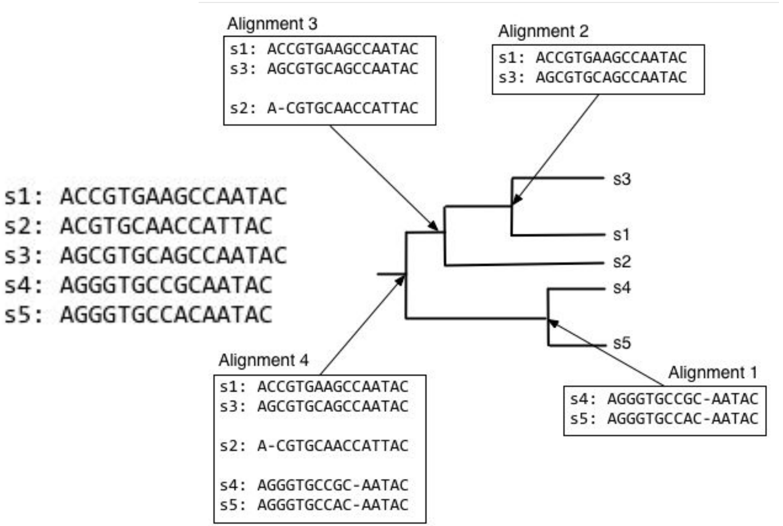
4. Evolutionary conservation || sequence alignment

4. Homology || Sequence alignment rationale

A **sequence alignment** is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences

Main idea:

- Score positively the matches, penalizing mismatches and/or gaps
- Residues (aa) relevant for a function are evolutionary "conserved", for example:
  - Promoters of housekeeping genes (essential for cell maintenance processes)
  - Protein domains important for a function are generally conserved
    - Zinc fingers, Disulfide bonds
    - Phosphorylation-related domains
- Alignments can be used to reconstruct the **phylogeny** of a set of species (evolutionary tree)



```
s1: ACCGTGAAGCCAATAC
s2: ACGTGCAACCATTAC
s3: AGCGTGCAGCCAATAC
s4: AGGGTGCCGCAATAC
s5: AGGGTGCCACAATAC
```

Alignment 3

```
s1: ACCGTGAAGCCAATAC
s3: AGCGTGCAGCCAATAC

s2: A-CGTGCAACCATTAC
```

Alignment 2

```
s1: ACCGTGAAGCCAATAC
s3: AGCGTGCAGCCAATAC
```

Alignment 4

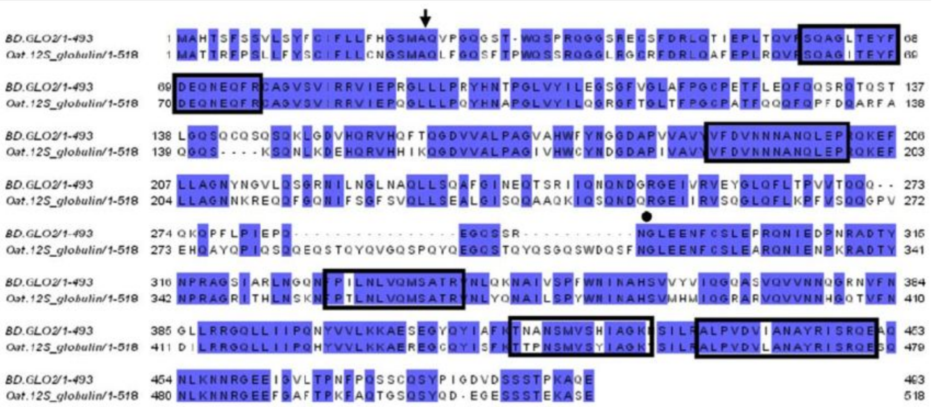
```
s1: ACCGTGAAGCCAATAC
s3: AGCGTGCAGCCAATAC

s2: A-CGTGCAACCATTAC

s4: AGGGTGCCGC-AATAC
s5: AGGGTGCCAC-AATAC
```

Alignment 1

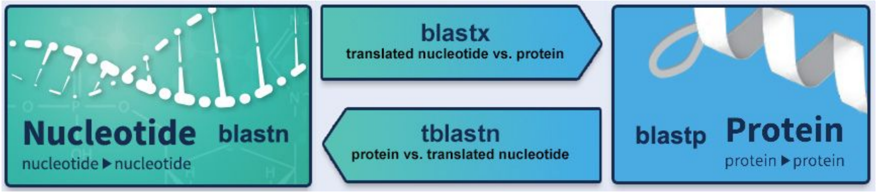
```
s4: AGGGTGCCGC-AATAC
s5: AGGGTGCCAC-AATAC
```



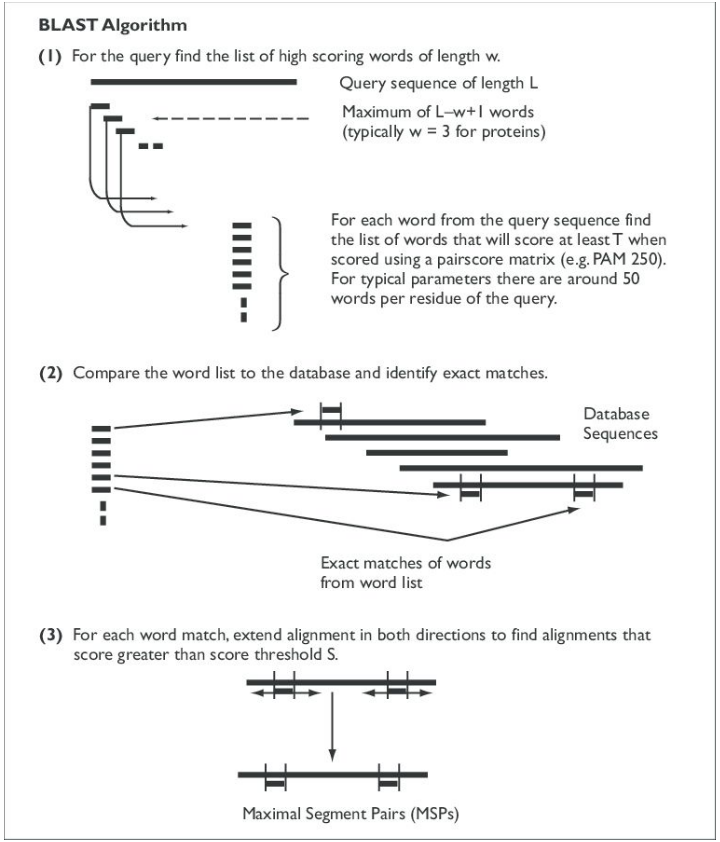
4. Homology|| BLAST algorithm

- Alignment of a sequence against annotated sequences databases
  - **Same sequence = same structure = same function**

BLAST web



- Importance of the DB used:
  - RefSeq
  - PDB
  - Curated databases (e.g., PlasticDB, etc)
  - OMD, motus-db
  - etc



BLAST Algorithm

(1) For the query find the list of high scoring words of length w.

Query sequence of length L

Maximum of L–w+1 words (typically w = 3 for proteins)

For each word from the query sequence find the list of words that will score at least T when scored using a pairscore matrix (e.g. PAM 250). For typical parameters there are around 50 words per residue of the query.

(2) Compare the word list to the database and identify exact matches.

Database Sequences

Exact matches of words from word list

(3) For each word match, extend alignment in both directions to find alignments that score greater than score threshold S.

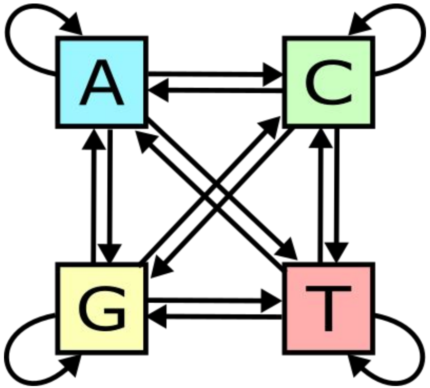
Maximal Segment Pairs (MSPs)

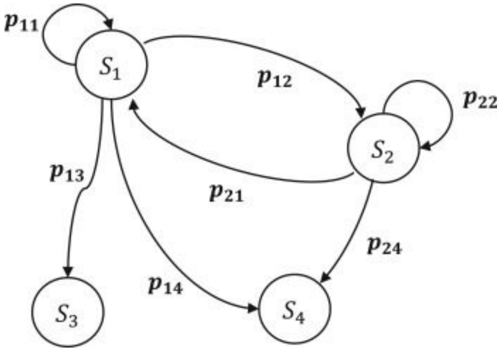
5. Evolutionary conservation || HMMs

5. HMMs || applications in biology

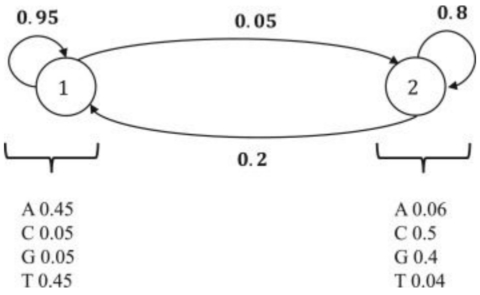
HMMs are often preferred over BLAST in biology because they better detect distant homologs, use profile-based matching to capture conserved sequence motifs, and offer greater accuracy in functional annotation through probabilistic modeling.

- In HMMs, the "hidden" states represent unknown biological properties (e.g., functional domains in a protein sequence), while the "observations" are the actual sequence elements (like nucleotide or amino acid residues). The model probabilistically associates these observable elements with specific hidden states, helping predict or annotate parts of the sequence.
- Aid in tasks like gene prediction, protein family classification, and functional annotation.
- Tools like HMMER use HMMs to scan large databases, identifying sequences that match known biological profiles, even with slight variations.



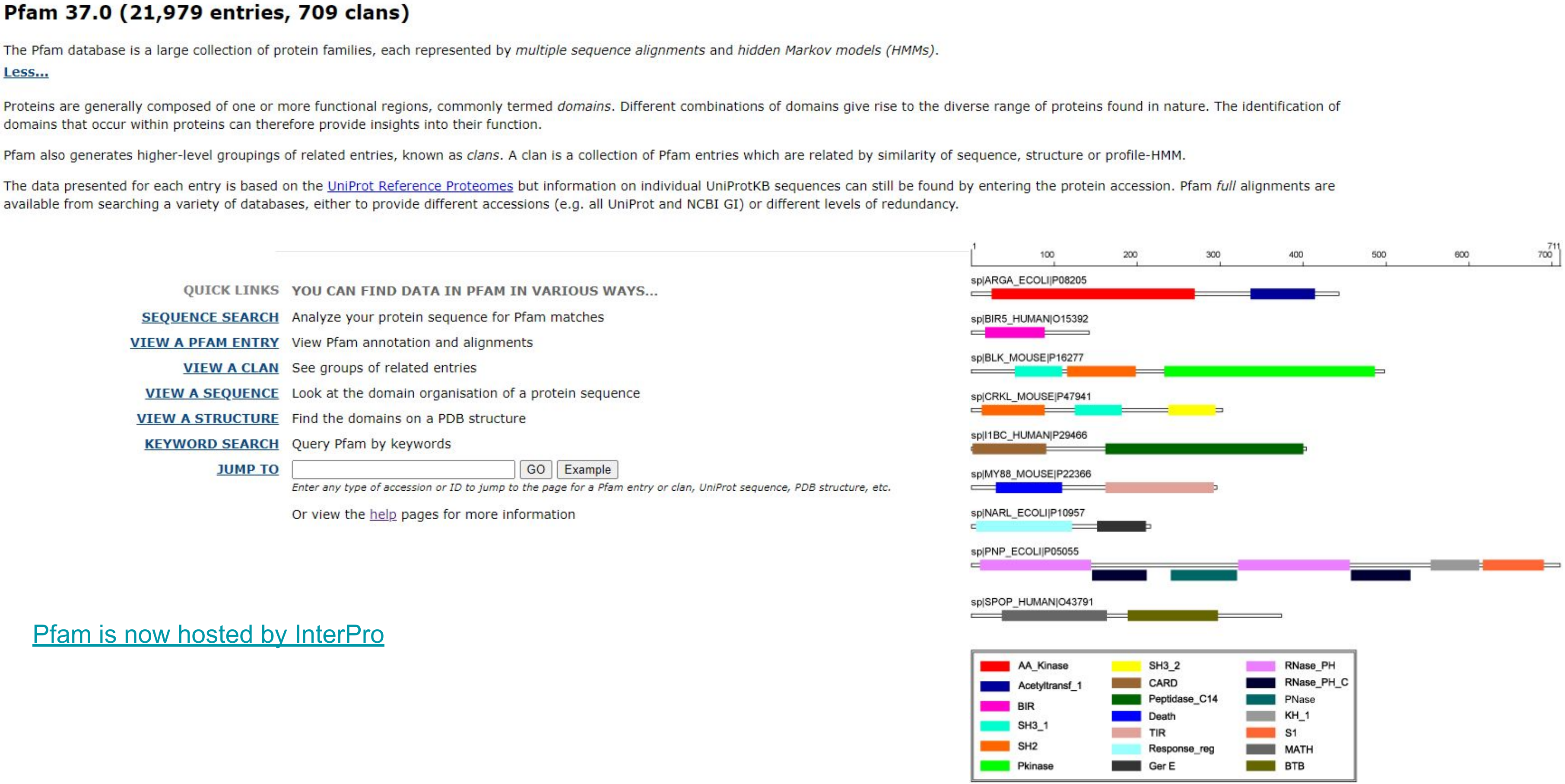


transitions



Transitions + emissions

5. HMMs | | Pfam



Pfam 37.0 (21,979 entries, 709 clans)

The Pfam database is a large collection of protein families, each represented by *multiple sequence alignments* and *hidden Markov models (HMMs)*.
Less...

Proteins are generally composed of one or more functional regions, commonly termed *domains*. Different combinations of domains give rise to the diverse range of proteins found in nature. The identification of domains that occur within proteins can therefore provide insights into their function.

Pfam also generates higher-level groupings of related entries, known as *clans*. A clan is a collection of Pfam entries which are related by similarity of sequence, structure or profile-HMM.

The data presented for each entry is based on the UniProt Reference Proteomes but information on individual UniProtKB sequences can still be found by entering the protein accession. Pfam *full* alignments are available from searching a variety of databases, either to provide different accessions (e.g. all UniProt and NCBI GI) or different levels of redundancy.

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM ENTRY	View Pfam annotation and alignments
VIEW A CLAN	See groups of related entries
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	GO Example *Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.*

Or view the help pages for more information

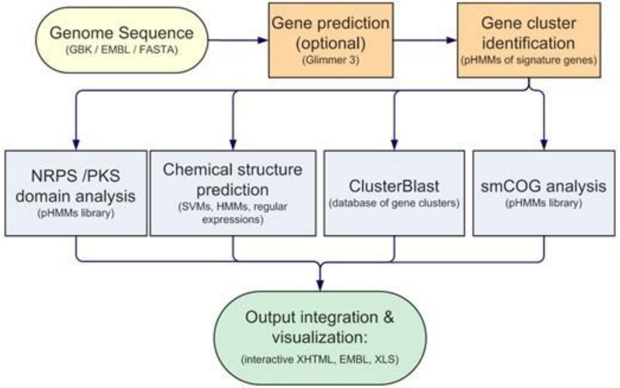
Pfam is now hosted by InterPro

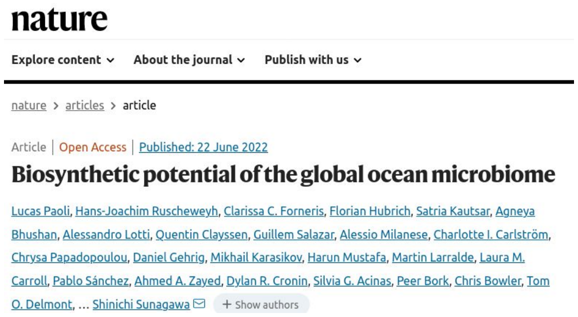
5. Advance models || antiSMASH

Metabolic gene clusters or **biosynthetic gene clusters (BGCs)** are tightly linked sets of (mostly) non-homologous genes participating in a **common, discrete metabolic pathway or biological process**.

- their expression is often coregulated (same operon, same TFs, etc.)
- Other factors such as cluster completion needs to be considered

anti
SMASH


Genome Sequence
(GBK / EMBL / FASTA)
Gene prediction
(optional)
(Glimmer 3)
Gene cluster identification
(pHMMs of signature genes)
NRPS /PKS domain analysis
(pHMMs library)
Chemical structure prediction
(SVMs, HMMs, regular expressions)
ClusterBlast
(database of gene clusters)
smCOG analysis
(pHMMs library)
Output integration & visualization:
(interactive XHTML, EMBL, XLS)



Article | Open Access | Published: 22 June 2022

Biosynthetic potential of the global ocean microbiome

Lucas Paoli, Hans-Joachim Ruscheweyh, Clarissa C. Forneris, Florian Hubrich, Satria Kautsar, Agneya Bhushan, Alessandro Lotti, Quentin Clayssen, Guillem Salazar, Alessio Milanese, Charlotte I. Carlström, Chrysa Papadopoulou, Daniel Gehrig, Mikhail Karasikov, Harun Mustafa, Martin Larralde, Laura M. Carroll, Pablo Sánchez, Ahmed A. Zayed, Dylan R. Cronin, Silvia G. Acinas, Peer Bork, Chris Bowler, Tom O. Delmont, … Shinichi Sunagawa ✉ + Show authors

- **40,000 putative new BGCs**
- High **discover potential**
  - New drugs
  - Novel biotechnological applications
  - New biological paradigms
  - etc.

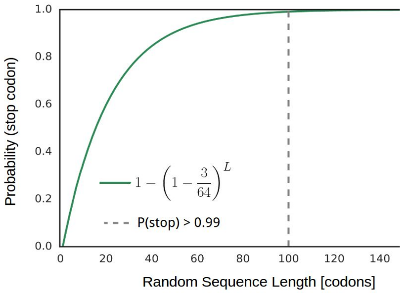
TAOJ22-1_SAMN27365860_MAG_00000152 - 55 region(s) - antiSMASH results

6. Sequence feature-based annotation || SEPs, AMPs and signal peptides

6. Sequence feature-based || SEPs

Gene annotation tools initial step: Open Reading Frame scanning

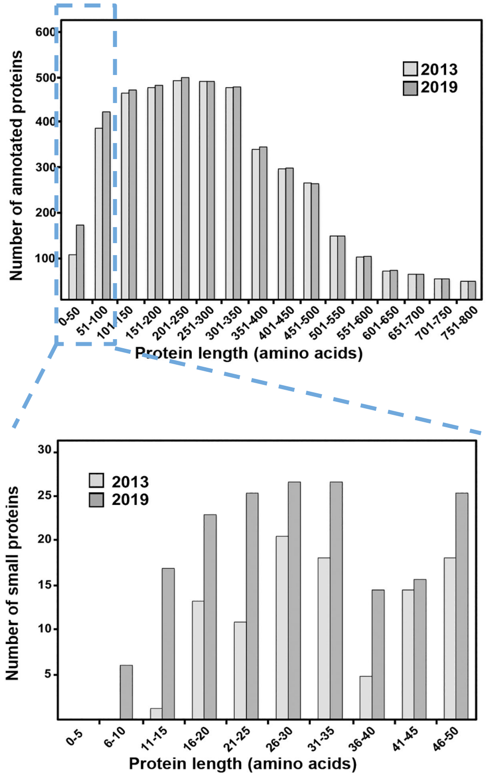
- Arbitrary threshold in 100 aa
  - Large % of small-ORFs will be non-coding



- small ORF-Encoded Proteins (SEPs, ≤100 aa)

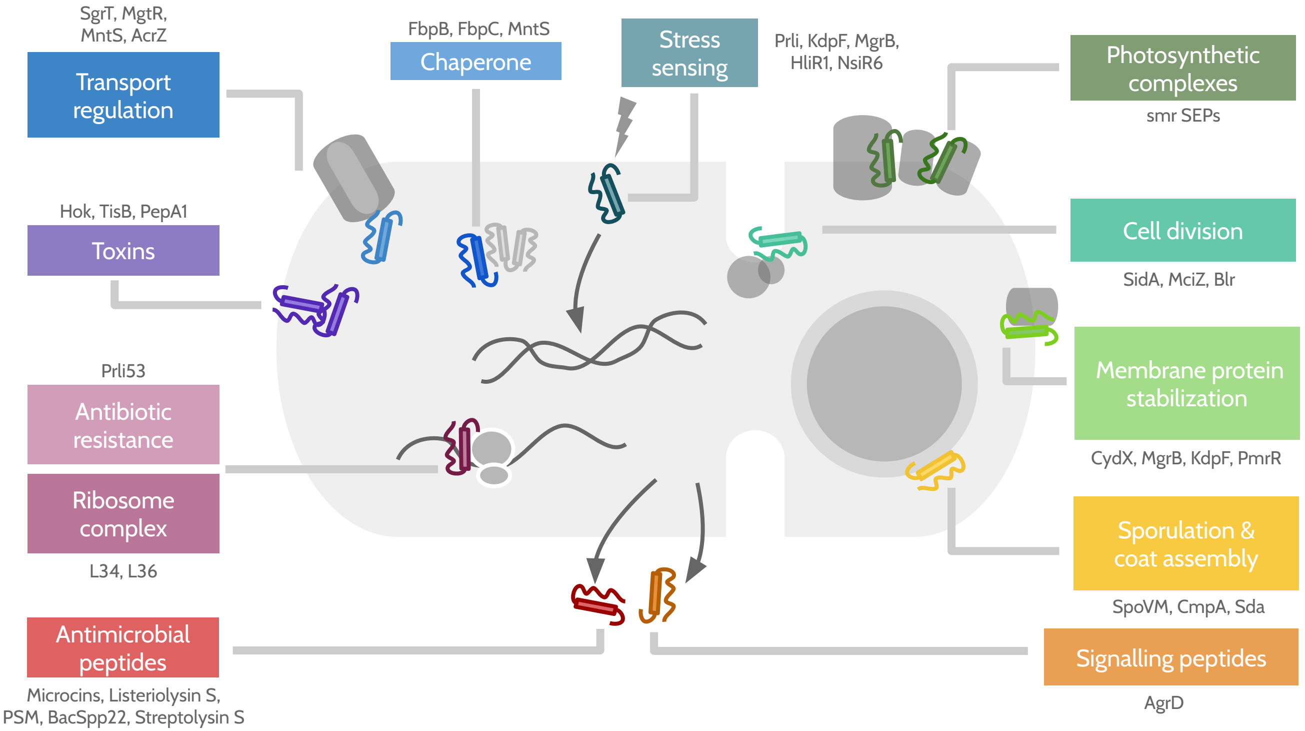
- Experimental approaches overlook SEPs
  - Under-representation in databases
  - Discovered by serendipity in screening assays
    - Epitope/Tag approaches → case by case Van Orsdel, CE. *et al.* (Proteomics, 2018)
    - Transcriptomics / RiboSeq Weaver, J. *et al.* (mBIO, 2019)
      - mRNAs translation but no frame information
    - Mass Spectrometry as top contributor Ahrens, CH. *et al.* (J. Bacteriol. Res., 2022)

Functional? Translational noise?



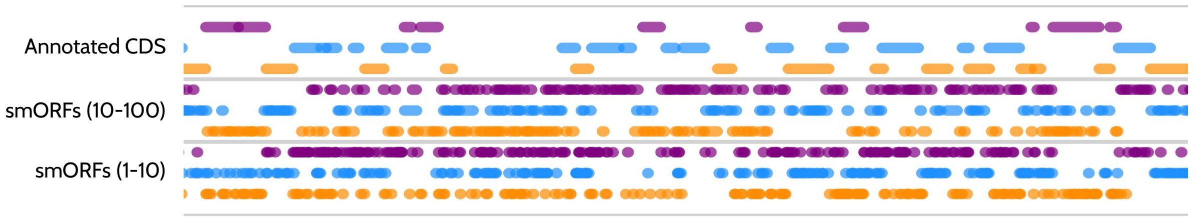
Total number of new SEPs doubled number in *E. coli.* Adapted from Hemm MR, *et al.* (EcoSal Plus 2020)

6. Sequence feature-based || SEPs

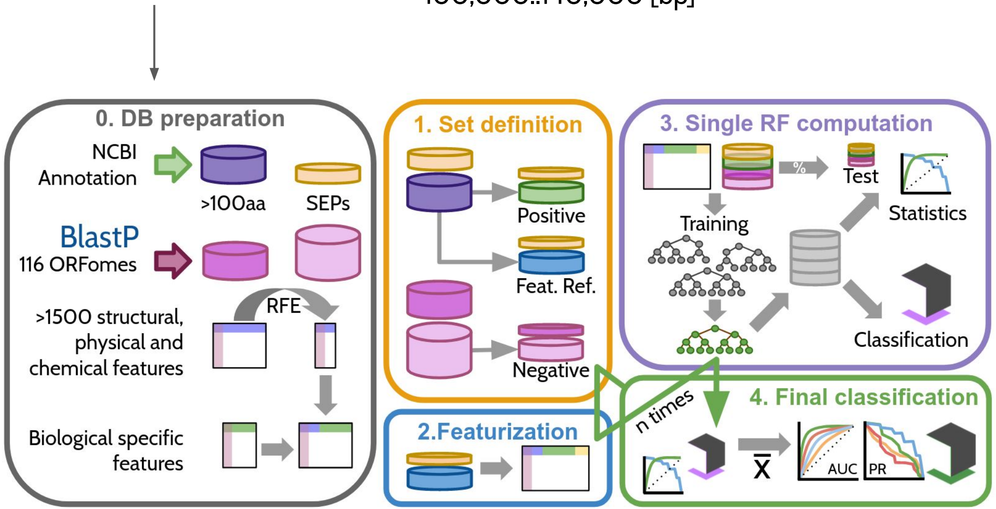


From 'Development of computational and experimental tools for the identification of small proteins in bacterial genomes' - S Miravet-Verde (2021) https://www.tdx.cat/handle/10803/671772

6. Sequence feature-based || SEPs



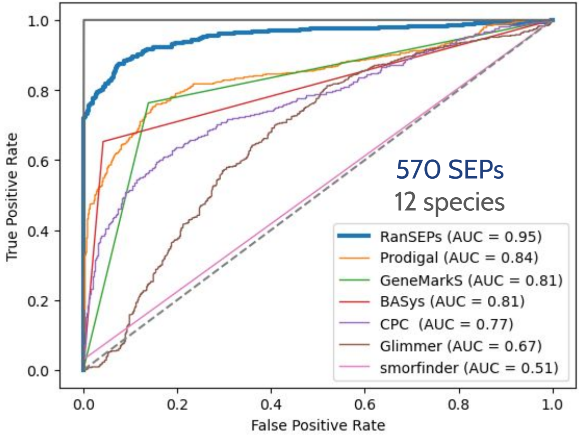
~5% of *M. pneumoniae* Genome
100,000..140,000 [bp]



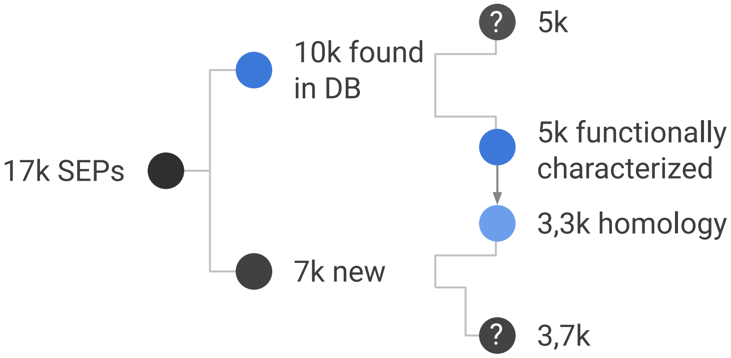
Miravet-Verde S; et al., 2019. *Unraveling the hidden universe of small proteins in bacterial genomes.* Mol Syst Biol 15(2):e8290

6. Sequence feature-based || SEPs

- **Findings**
  - + 4k in human-gut Sberro H, et al. (Cell, 2019)
  - + 40k in phages Fremin BJ, et al. (Cell Rep., 2022)
  - - limited discovery
- 50% no associated function



570 SEPs
12 species



17k SEPs

10k found in DB

5k

5k functionally characterized

3,3k homology

7k new

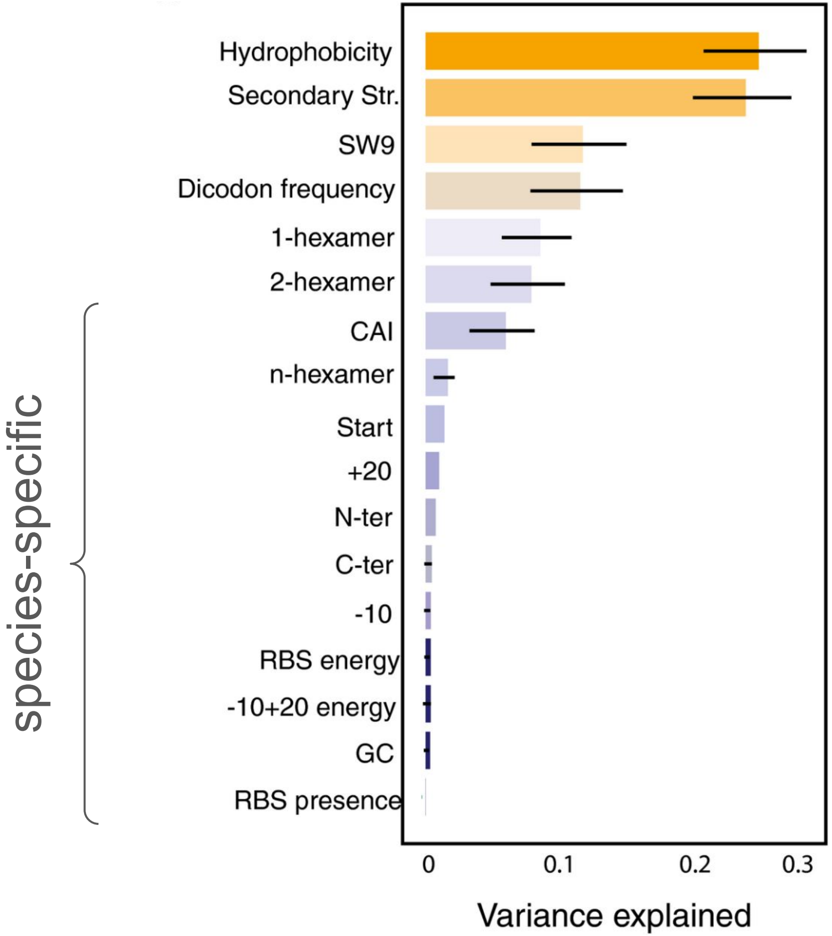
3,7k

Secreted/exposed SEPs are enriched:

- 25% AMPs
- 10% signal peptides
- 15% transmembrane motif

6. Sequence feature-based || SEPs

- Can we predict in an species-agnostic manner?
  - ~80% of the variance explained by species-independent features
  - Testing:
    - Training without species consideration
    - Evaluate prediction accuracy
    - Consider using pre-computed databases



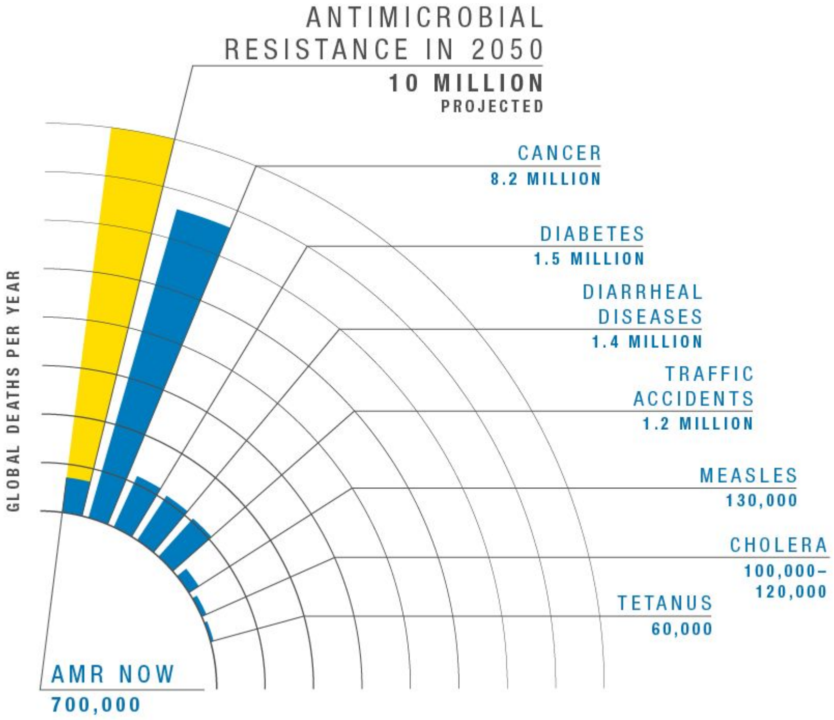
6. Sequence feature-based || AMPs

We need to broaden the search for
novel bioactive compounds

new environments

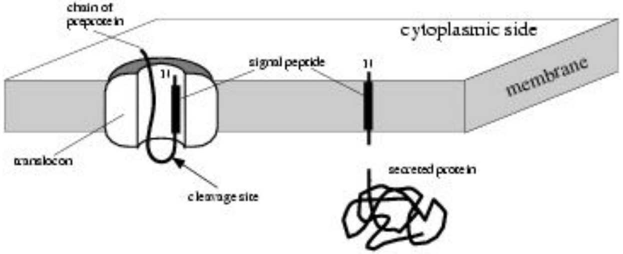
new functions

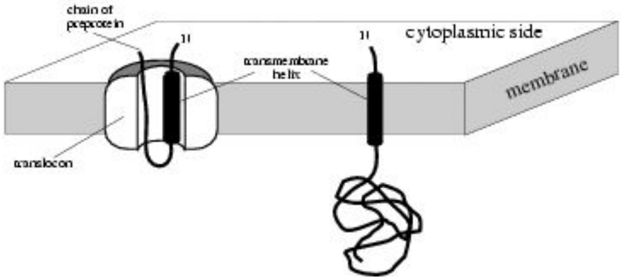
- Traditional methods to identify novel bioactive compounds (*e.g.*, antimicrobials):
  - economically and timewise expensive
  - cultivability required
  - P(re-identification) > P(discovery)

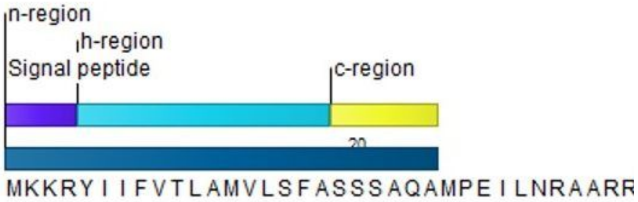


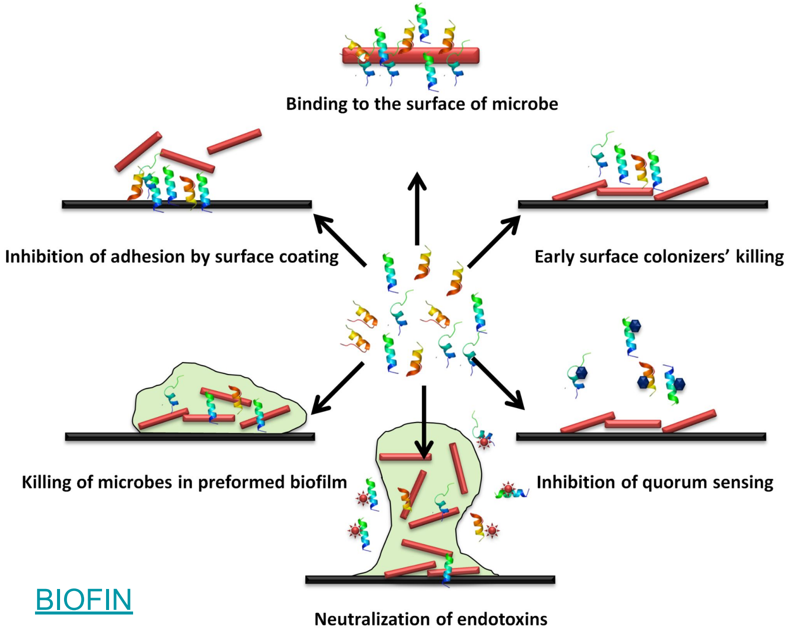
"Environmental genomics-mediated discovery"

6. Sequence feature-based || Signal peptides



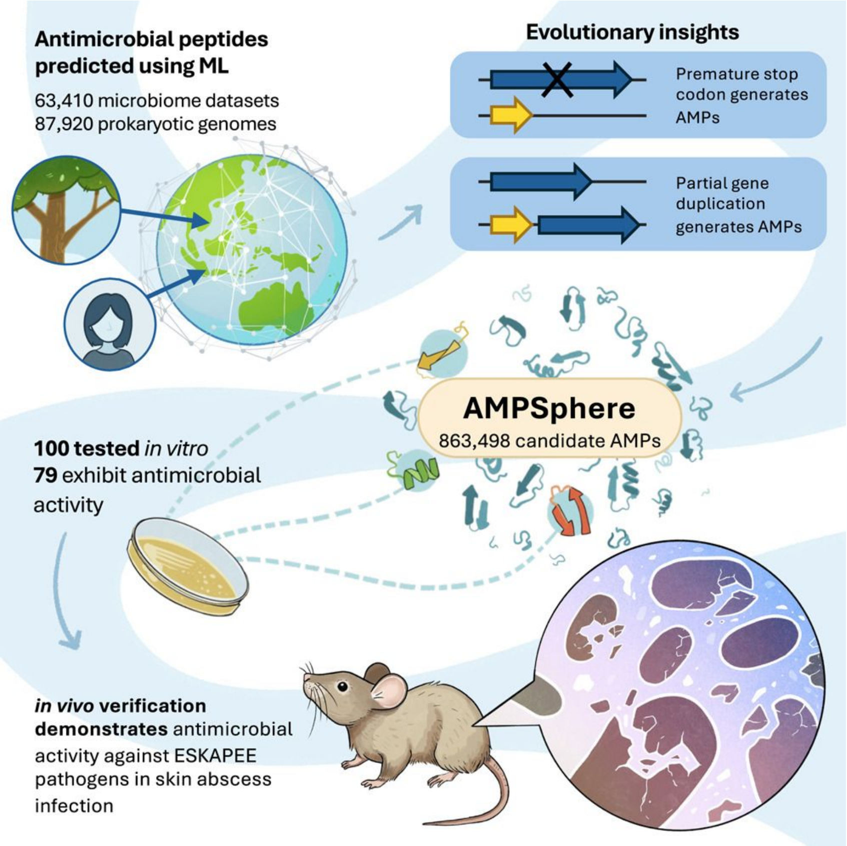





Binding to the surface of microbe
Inhibition of adhesion by surface coating
Early surface colonizers' killing
Killing of microbes in preformed biofilm
Inhibition of quorum sensing
Neutralization of endotoxins

BIOFIN

6. Sequence feature-based || AMPs



Antimicrobial peptides predicted using ML

63,410 microbiome datasets
87,920 prokaryotic genomes

Evolutionary insights

Premature stop codon generates AMPs

Partial gene duplication generates AMPs

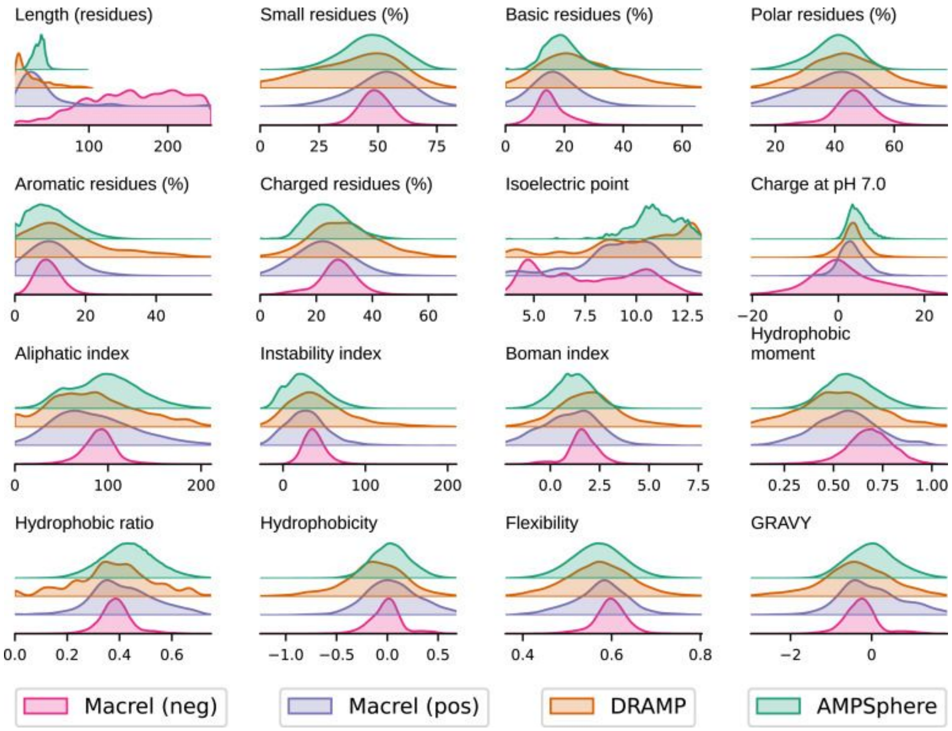
AMPSphere

863,498 candidate AMPs

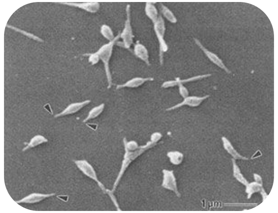
100 tested *in vitro*
79 exhibit antimicrobial activity

***in vivo* verification demonstrates** antimicrobial activity against ESKAPEE pathogens in skin abscess infection

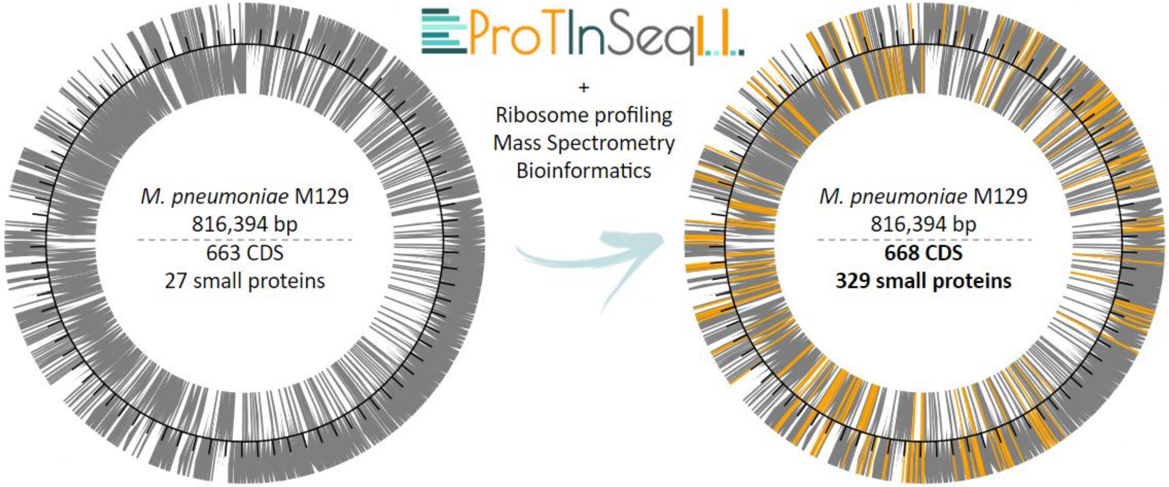
https://www.cell.com/cell/fulltext/S0092-8674%2824%2900522-1



Small proteins | | What are we missing?



- **Downsized genome** (816 kb) → **systems** and **synthetic biology** model
  - ○ ‘Simple’: 689 protein coding genes annotated (27 SEPs)
  - ○ Weak lung pathogen → biomedical/veterinary applications

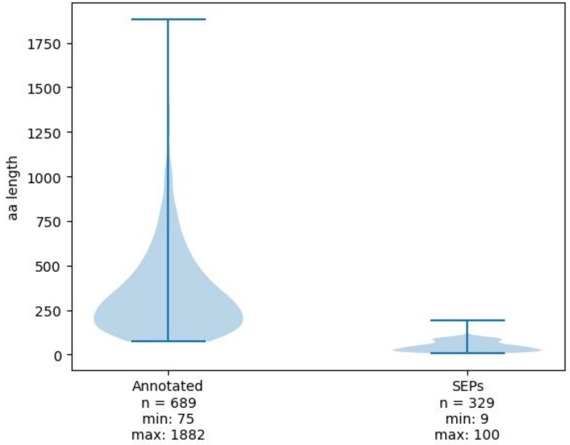


M. pneumoniae M129
816,394 bp
663 CDS
27 small proteins

ProTInSeq

+
Ribosome profiling
Mass Spectrometry
Bioinformatics

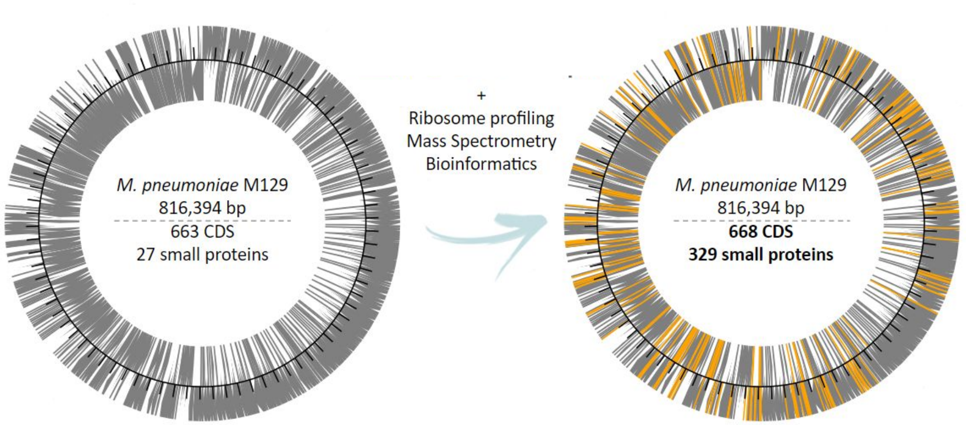
M. pneumoniae M129
816,394 bp
668 CDS
329 small proteins



‘ProTInSeq: TnSeq applied to protein detection, quantification and functional studies’ - S Miravet-Verde, et al. Nat Comm 2024

Discussion || Open questions

- Coding potential
  - M. pneumoniae 690 → 997 genes
  - E. coli 4,000 → ?
  - Human 20,000 → ?



- **Functional** characterization is still a **problem,** still great potential for therapeutic/biotechnological applications:
  - **~50%** of the identified SEPs are hypothetical or unassigned-function proteins
  - **Secreted** and **membrane** located SEPs
    - Similar results in **metagenomic** studies Sberro H, et al. (Cell, 2019) :
      - >4,000 conserved SEPs families
        - 30% predicted to be secreted and/or transmembrane
      - However, **other ecological niches yet to be addressed** → **ocean?**

7. Protein structures || Latest advances

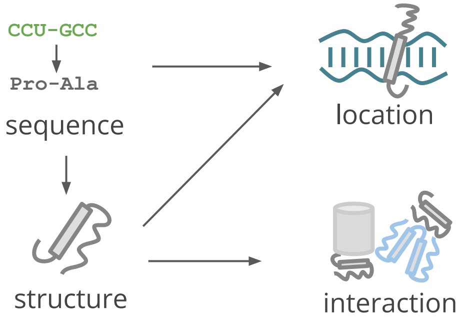
7. Structural biology || Mechanisms from sequence and structures

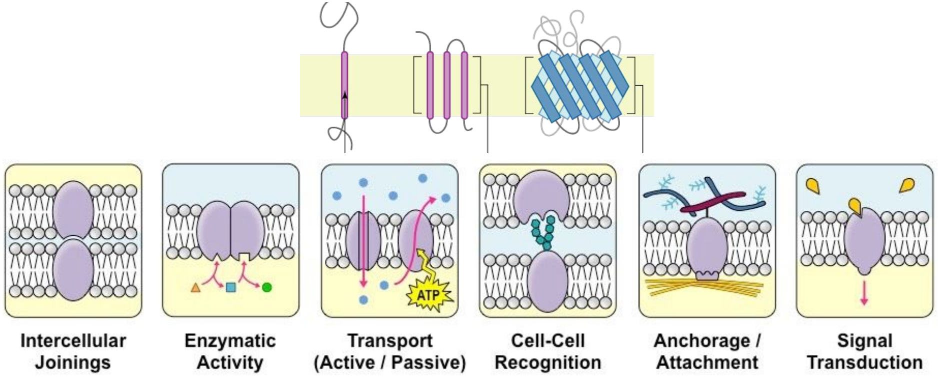
Proteins function by **interacting** with other molecules (DNA, RNA, proteins and metabolites)

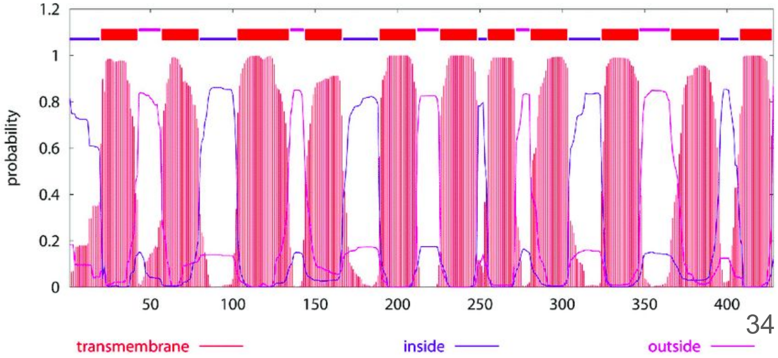
- **Structural** roles (e.g. collagen)
- **Globular** proteins → they are soluble in water and function in and out the cell
  - Catalytic roles → enzymes
- **Membrane**-associated proteins
  - cell communication and transport
  - protein channels
- **Secreted** proteins to interact with other members in an ecosystem:
  - Signal peptides → communication
  - Antimicrobial peptides → competition

Each of these will present **specific protein domains** and amino acid compositions

- There are databases to find these motifs in new sequences (PFAM, Uniprot, etc.)
- There are software tools to predict localization and transmembrane domains:

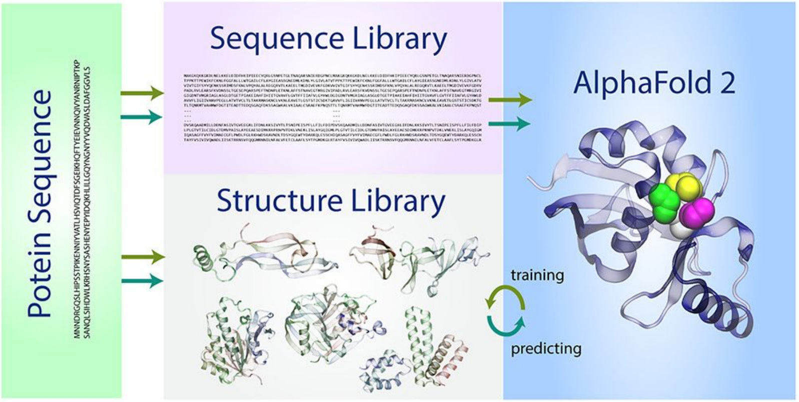






7. Structural biology || Predicting structures directly from sequence

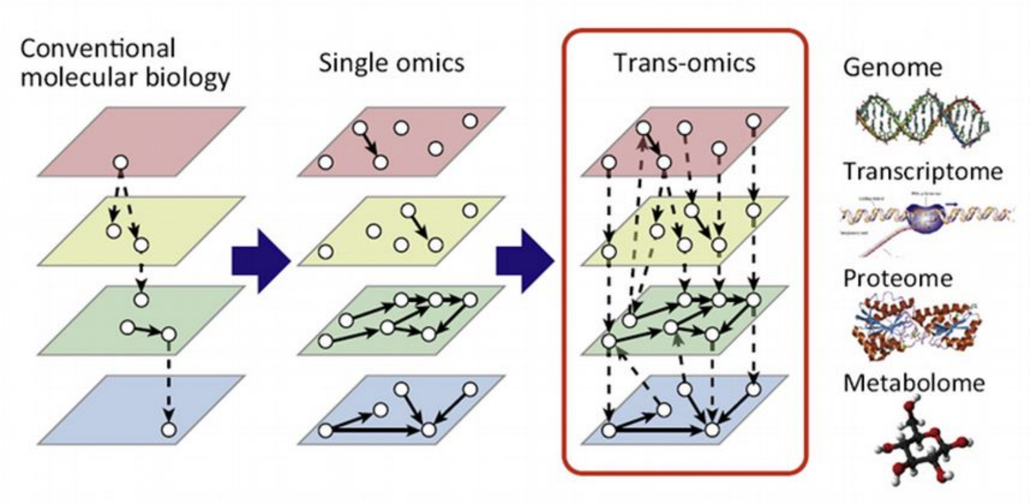
- Alignment + Artificial intelligence models trained with known structures allow now to **predict the structure of proteins**


Potein Sequence
MNNDRGQSLHIPSSTPIKENNIYVATLHSVIQTDFSGEIKHQFTYEIEVNNQIVYANRNIPTKP
SANQLSIHDWLKRHSNYSASHENYEPYIDQKHLILLGQYNGNYYVQDVASLDAFGGVLS
Sequence Library
Structure Library
AlphaFold 2
training
predicting

AlphaFold 2 and 3: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function | Journal of Chemical Information and Modeling (acs.org) [https://pubs.acs.org/doi/10.1021/acs.jcim.1c01114]

8. Closing remarks

6. Closing remarks | | Integrative genomics



Methodology

- NGS

- RNASeq (NGS)
- microarray

- Mass spectrometry

- Mass spectrometry
- NMR

All this approaches tend to work with databases of already **known genes**

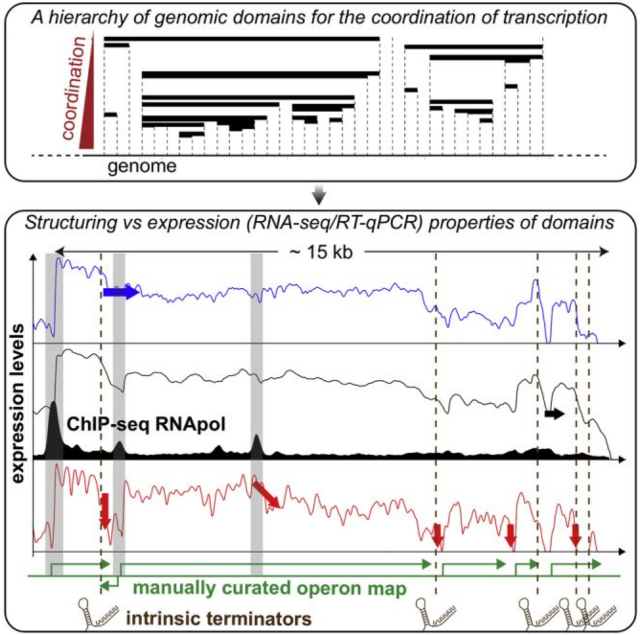
- A big fraction of the genes considered have no function associated → **growing knowledge**
- **Genome** exploration and comparative are grounding sources of **biological information**
  - Can be **extended** and **integrated** with other **omics** studies
- Tons of data (**big data**) → **computers** are essential
- **Bioinformatics** provide the tools required to **evaluate** and **validate**
  - New algorithm approaches, such as using Artificial Intelligence, are providing new paradigms in the way we integrate and understand biological information
- **Researchers** are still in the only “machines” capables of **interpreting** this data

0. Extra material

3. *Ab initio* || Regulatory elements: Terminators

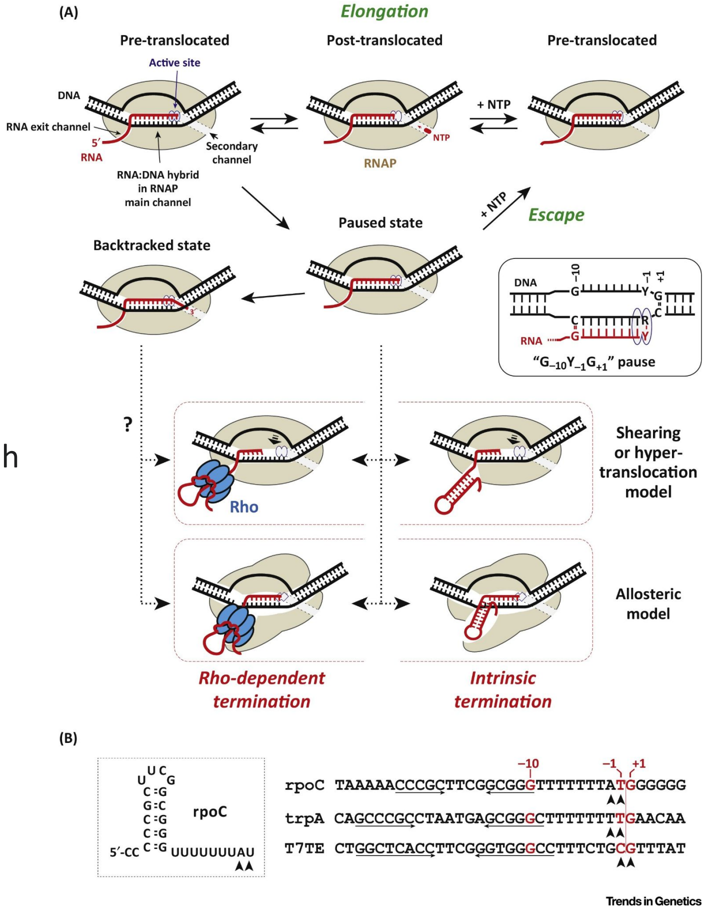
Transcriptional **termination** is associated to two types of processes:

- Factor-independent (also called intrinsic termination):
    - Relies on “**terminators**”, formed by a **secondary structure** in the transcribed RNA and a **poly-U** track
- Rho-dependent
    - Performed by the ***Rho* protein** which recognizes a GC-rich motif in the transcript



Terminators do not just finish transcription, they can regulate co-expression responding to external factors such as temperature (which affects RNA 2^{ndary} structures)

Insights into the Mechanisms of Basal Coordination of Transcription Using a Genome-Reduced Bacterium - Junier I. *et al.* (2016)



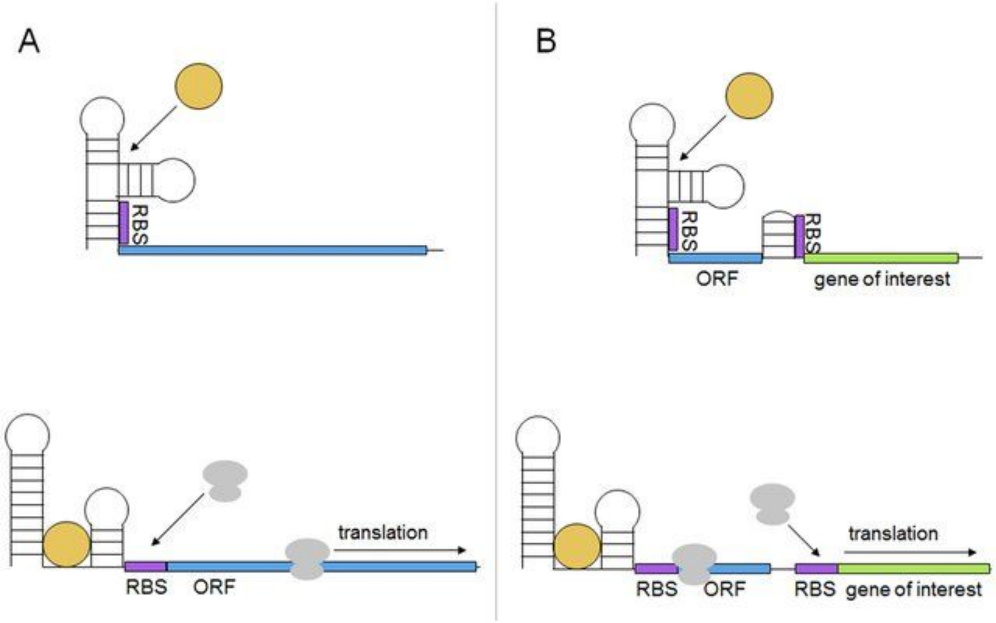
Transcription Termination - Porrua O. *et al.* (2016)

3. *Ab initio* || Regulatory elements: Ribosome Binding Sites

Ribosome binding sites (RBS) are in charge of recruiting ribosomes to start the translation of a messenger RNA (mRNA)

- They are found ~7 bp upstream a gene start codon (in the Untranslated Region [UTR] of mRNAs)

Additionally, they might be found associated to **Riboswitches**, RNA secondary structures that can interact with certain **metabolites** or **environmental conditions** (e.g. temperature) to hide/expose a RBS to control translation of a certain protein



RNA secondary structures related to terminators and Riboswitches can be predicted computationally:

ARNold

FINDING TERMINATORS

[Riboswitch Scanner (iiserkol.ac.in)](#)

[Riboswitch Finder (uni-wuerzburg.de)](#)