from genomics to functional biology understanding

Block Course Fall 2022
551-1119-00L
Microbial Community Genomics

Samuel Miravet-Verde
smiravet@ethz.ch

Guillem Salazar
guillems@ethz.ch

ETH zürich
D BIOL

9-Nov-22

Introduction || index

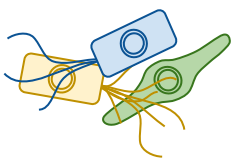
GOAL

Understand the **different features** that can be explored **from genomes** providing **mechanistic** and **functional** information about **biological systems**

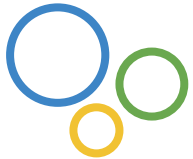
1. Introduction: the genomics rationale
2. Gene annotation
3. *Ab initio* gene annotation
    a. Sequence content
    b. Genetic elements
    c. Evaluation of sequence motifs
4. Evolutionary conservation: sequence alignment
5. From gene to function
6. Closing remarks

1. Introduction || The genomics rationale

What sequencing provides so far:



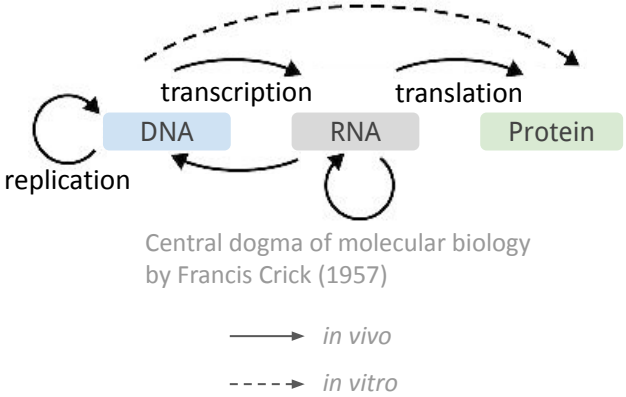
isolated cells or communities



genomic DNA sequences

???

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)

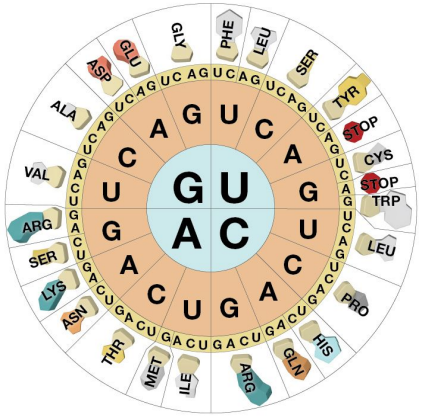
polymer	Deoxyribonucleic acids (DNA) - ACGT	Ribonucleic acids (RNA) - ACGU	20 amino acids
unit	gene	mRNA, tRNA, rRNA, ncRNA…	protein
set	genome	transcriptome	proteome
function	information	intermediary & regulation	structural & biochemical

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

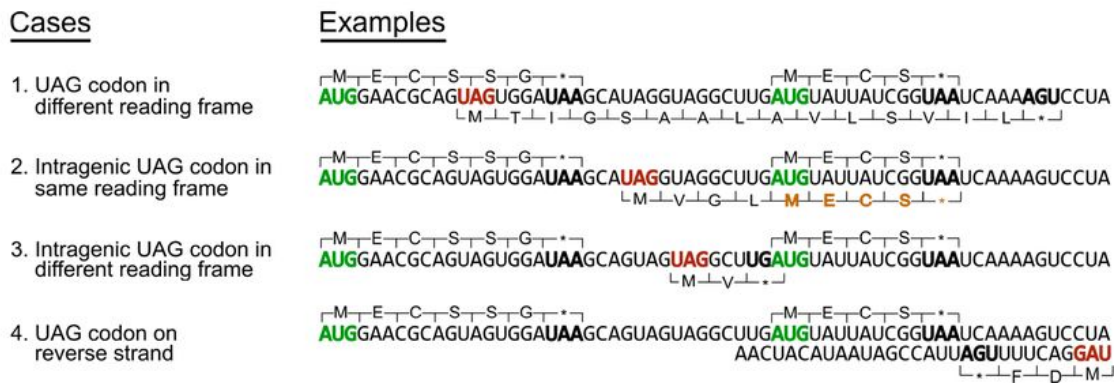
2. Gene annotation|| ORF scanning

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)



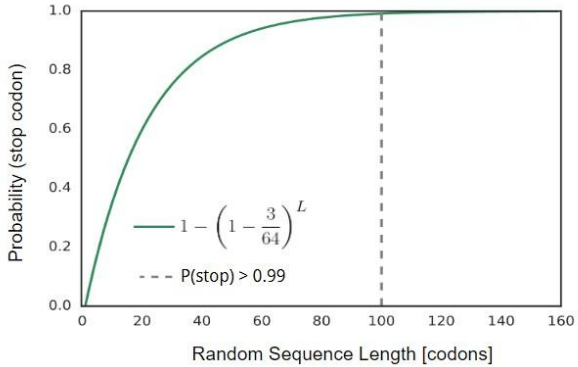
Genetic code in a codon table

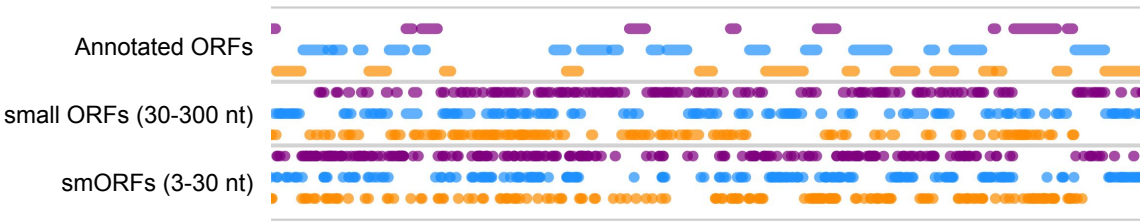


Complex for humans, very easy task for a computer

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

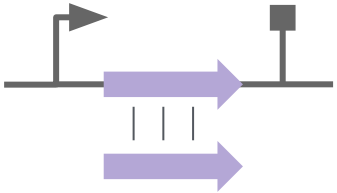
Ab initio:

- 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):

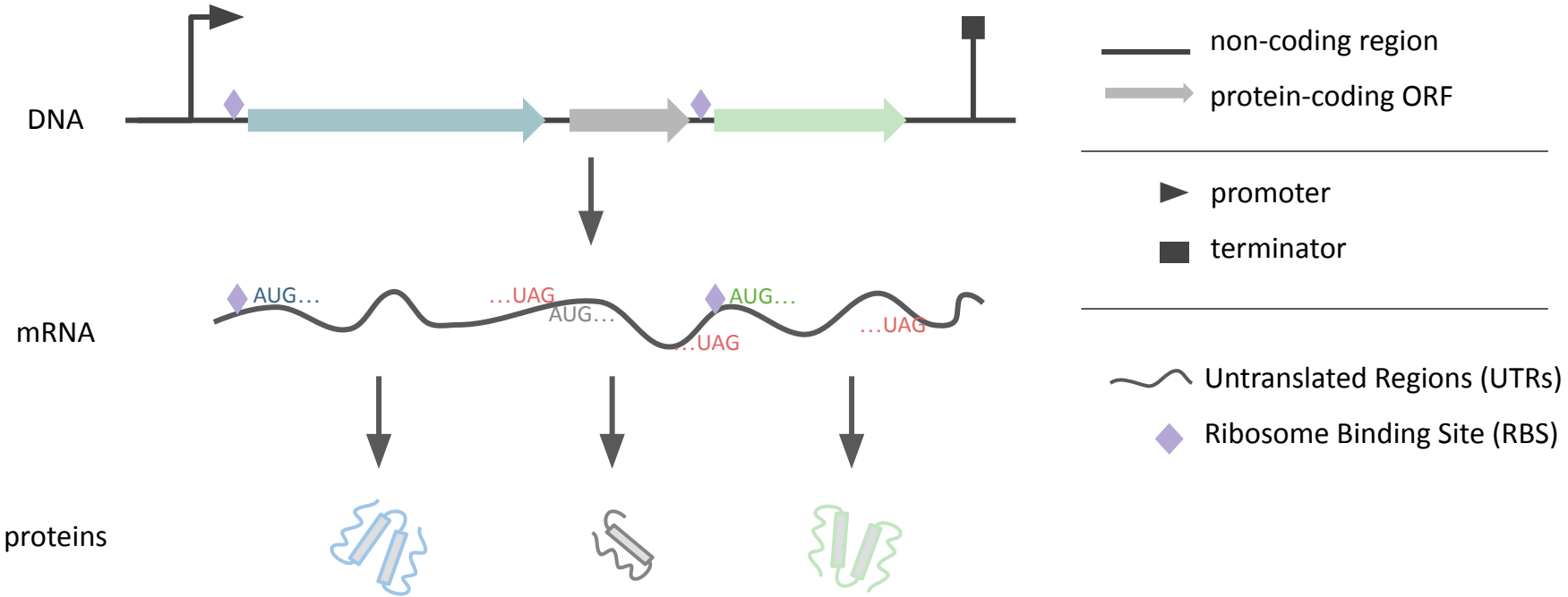


- Widely used in **genomic databases**
- NCBI Prokaryotic Genome Annotation Pipeline (PGAP) is the tool that runs when we submit a genome to NCBI

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”

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

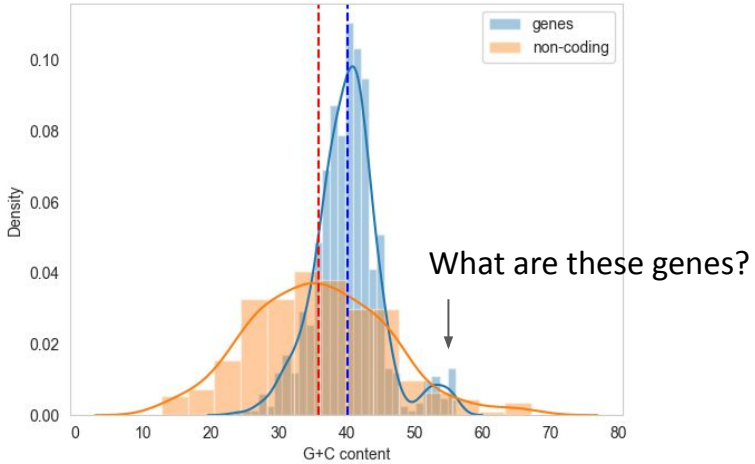


Sequence content and genetic features can all be explored at the DNA level and provide additional genetic insights

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



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

Other implications:

Not trivial to extrapolate mechanistic features…

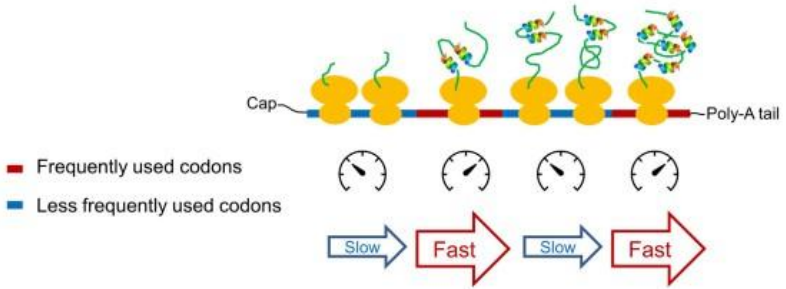
3. *Ab initio* || Sequence composition: Codon composition

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

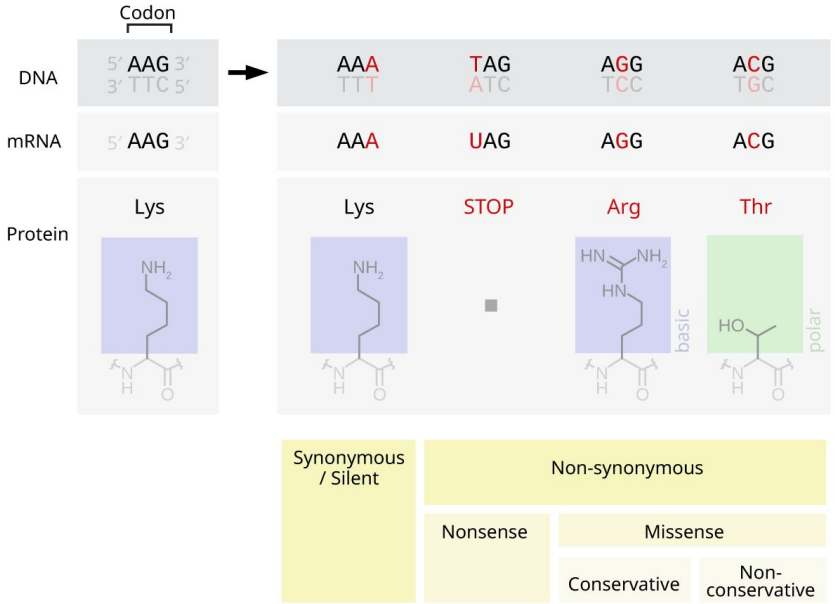
Arg codon frequencies in 4 model organisms

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)

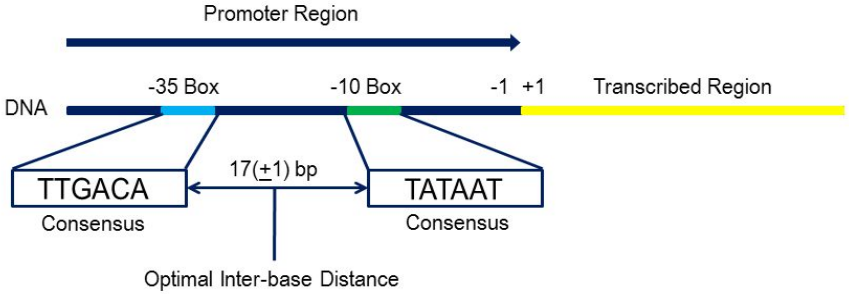
Mutation types:



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: Promoters

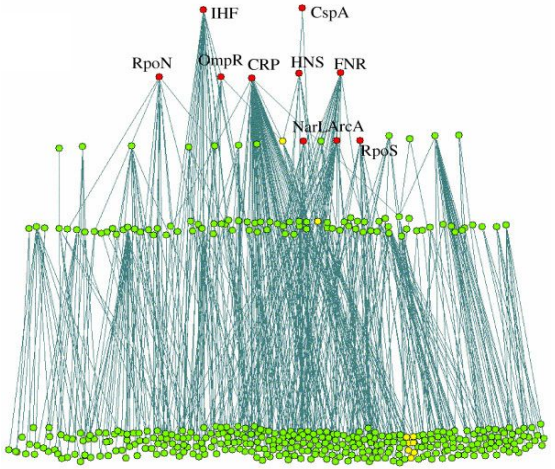
A **promoter** is a sequence of DNA to which proteins bind to initiate transcription of a single RNA transcript



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

- Genes expressed under the same promoter → **operon**
  - Coregulation 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*

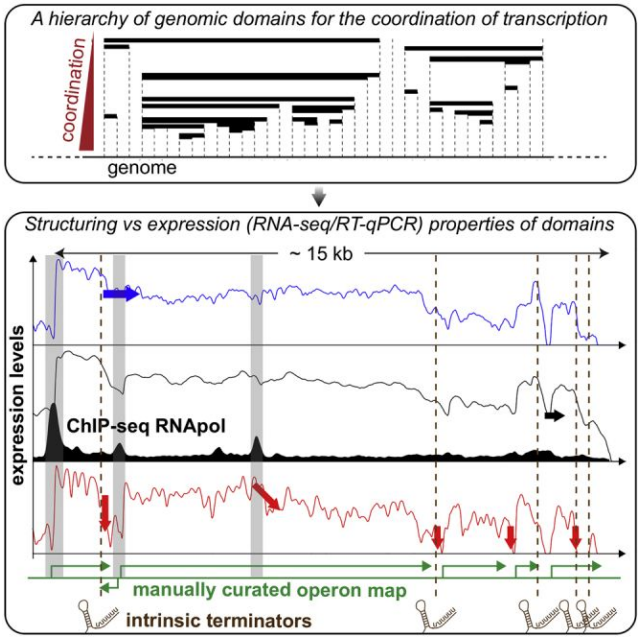
Certain TF are active under specific conditions (e.g., cold-shock, heat-shock, osmotic stress…)

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

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

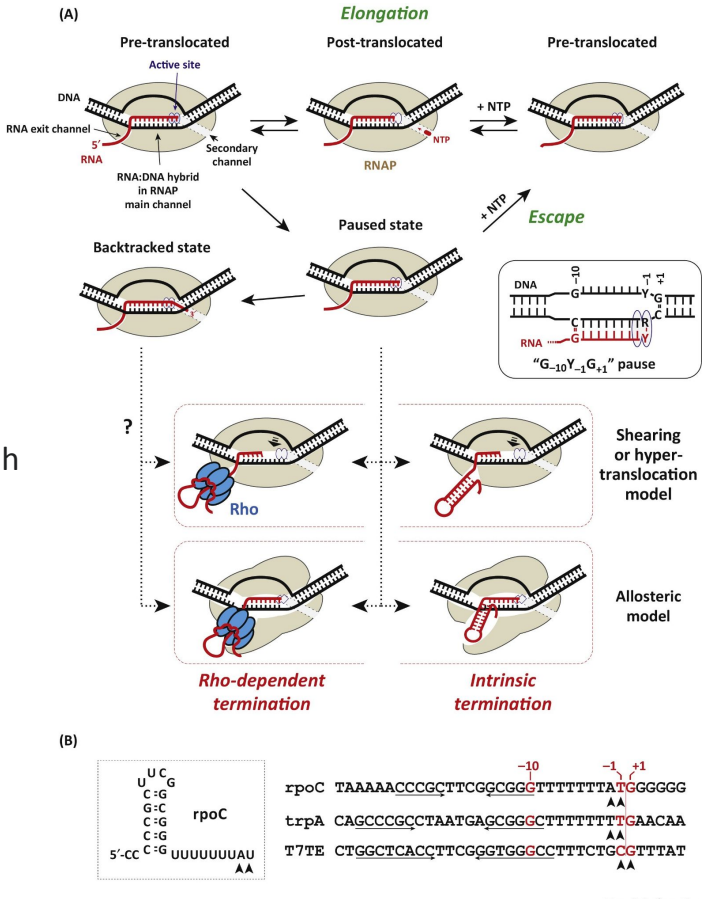
Transcriptional **termination** is associated to two types of processes:

- 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)



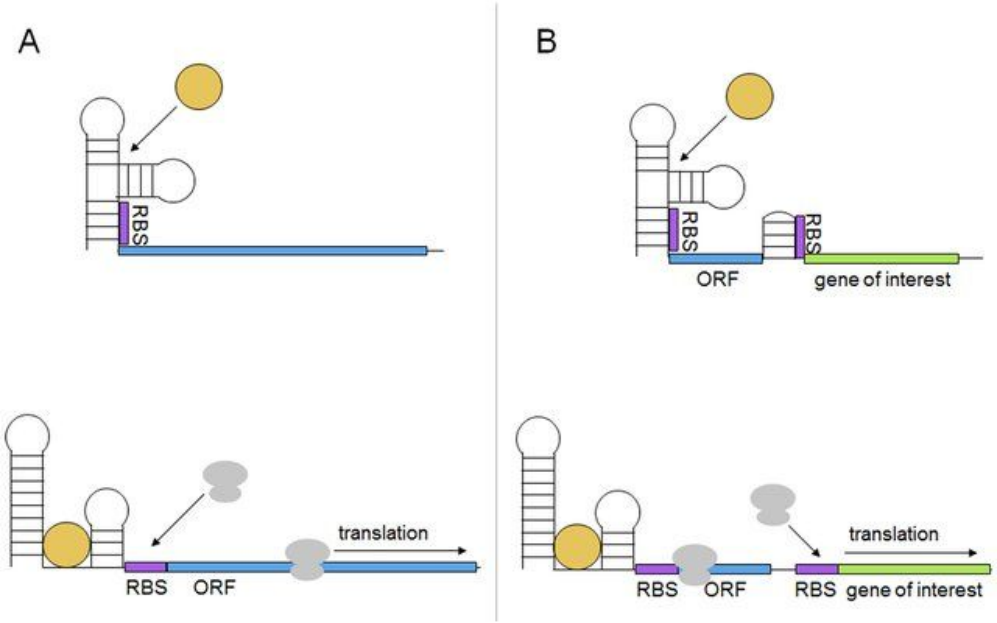
Transcription Termination - Porrua O. *et al.* (2016)

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)

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)

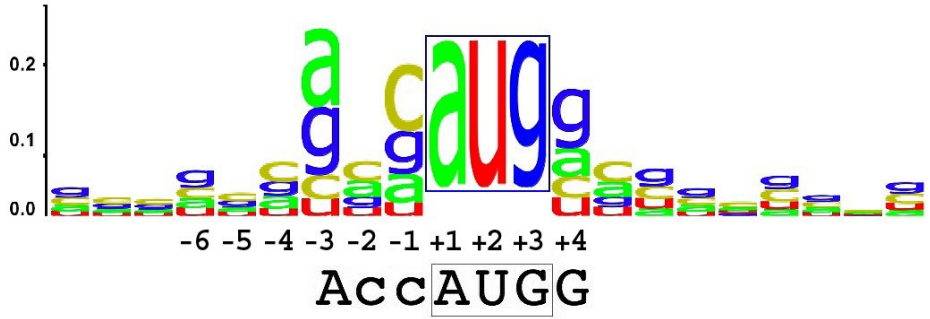
- These motifs may vary between species → evolution as driving force
- **Distance** between the regulatory motif and the regulated gene also matters

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

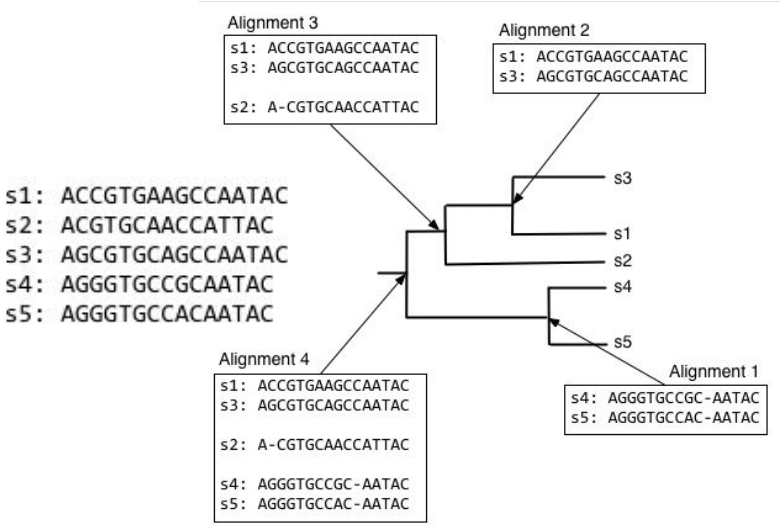


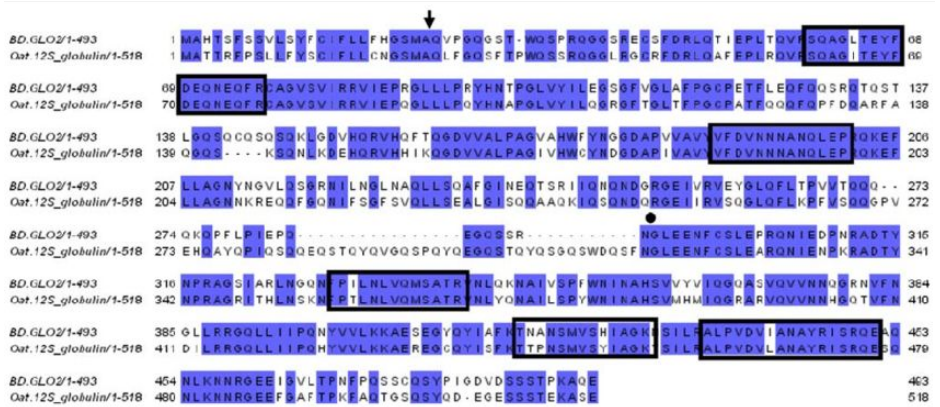
This same approach works with amino acid sequences

4. 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)

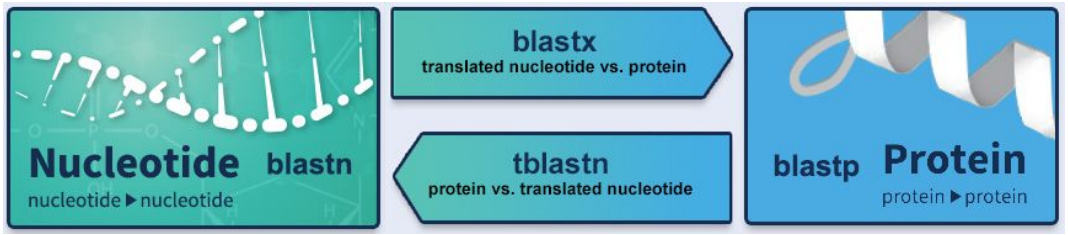




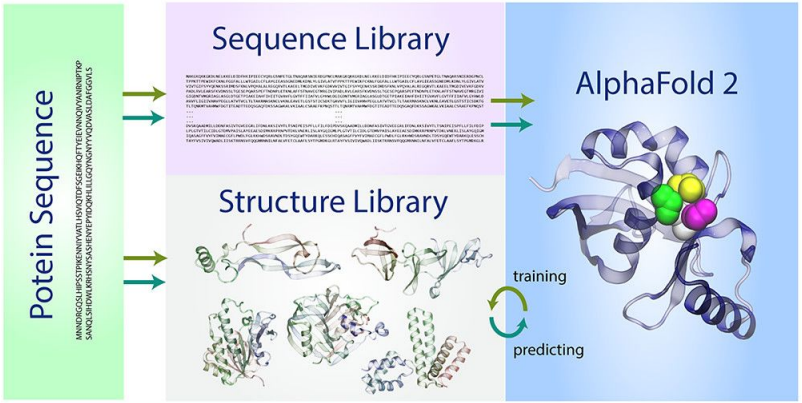
5. From gene to function || sequence alignment applications

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

BLAST web



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

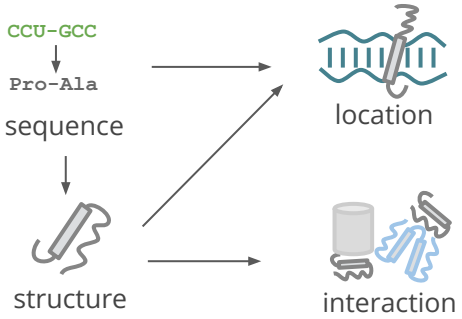


AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function | Journal of Chemical Information and Modeling (acs.org) [https://pubs.acs.org/doi/10.1021/acs.jcim.1c01114]

5. From gene to function || Mechanisms from sequence and structures

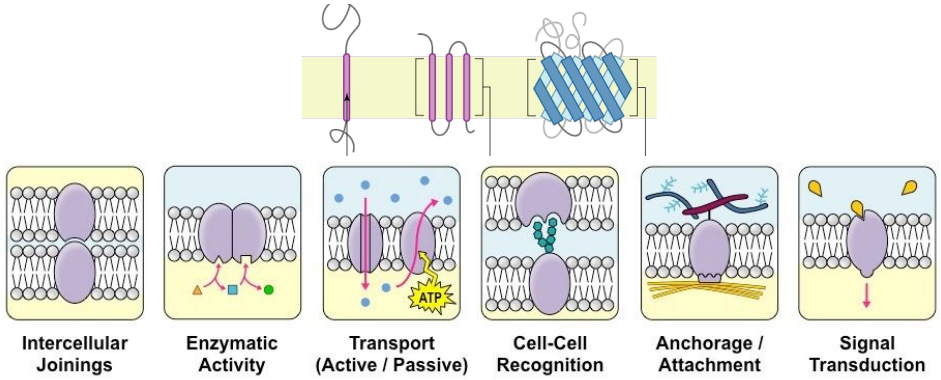
Proteins function by **interacting** with other molecules (DNA, RNA, proteins and metabolites)

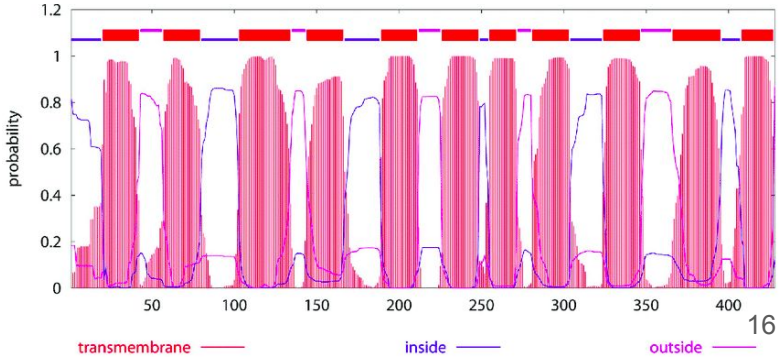
- **Structural** roles (e.g. collagen)
- **Globular** proteins → they are soluble in water and function in and out the cell
  - Catalytic roles → enzymes
- **Membrane**-associated proteins
  - cell communication and transport
  - protein channels
- **Secreted** proteins to interact with other members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

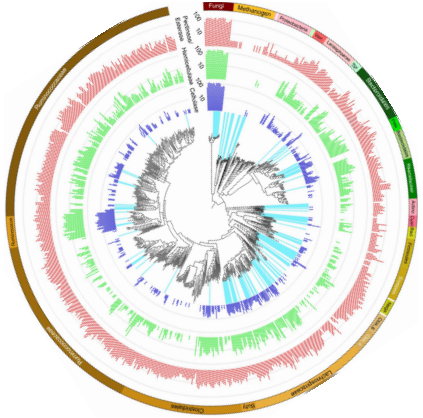


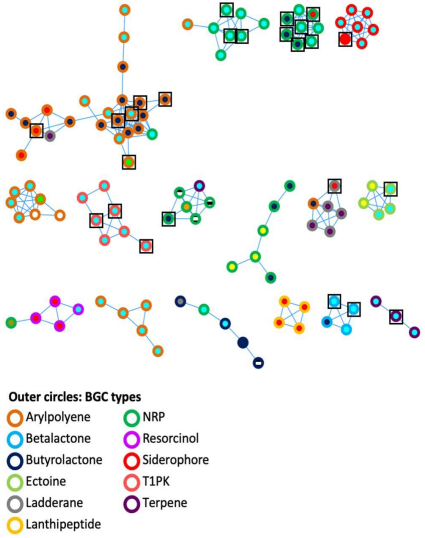


6. Closing remarks|| Back to genomic scale

Operational taxonomic unit or **OTU** is considered as the basic unit used in numerical taxonomy. These units may refer to an individual, species, genus, or class. OTUs are analytical units used in microbial ecology.

- Sequences can be clustered according to similarity (alignment).
  - The **16s ribosomal RNA** gene is commonly used to study the microbial community
  - **Single-copy genes** conserved at taxonomic levels can also be used to profile a community (e.g. **mOTUs**)





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.)


nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

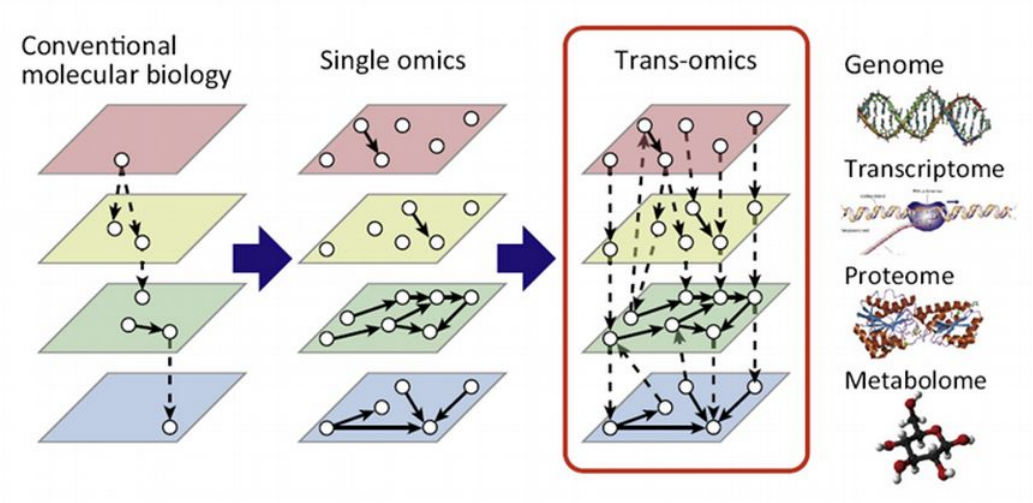
Article | Open Access | Published: 22 June 2022

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.
- Main tool: **antiSMASH**

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

from genomics to functional biology understanding

Block Course Fall 2022
551-1119-00L
Microbial Community Genomics

Samuel Miravet-Verde
smiravet@ethz.ch

Guillem Salazar
guillems@ethz.ch

ETH zürich
D BIOL

9-Nov-22