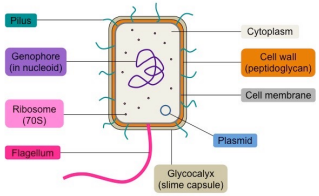
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM
Unidad Académica del Estado de Yucatán

Prediction of Transcription factors

Ernesto Pérez Rueda

ernesto.perez@iimas.unam.mx

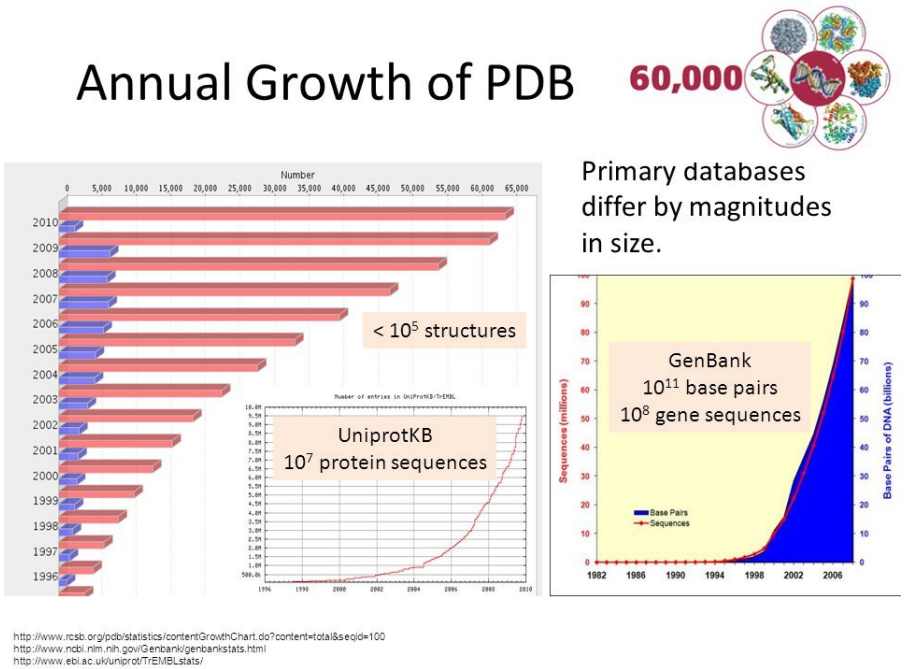


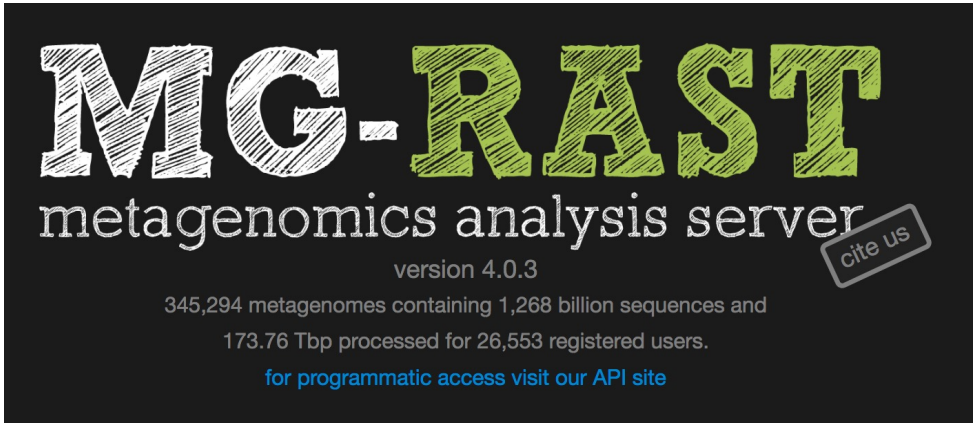
Bioinformatics, 2024

IIMAS - UNAM

http://openwetware.org/wiki/User:Ernesto_Perez-Rueda

The impact of genome sequencing projects





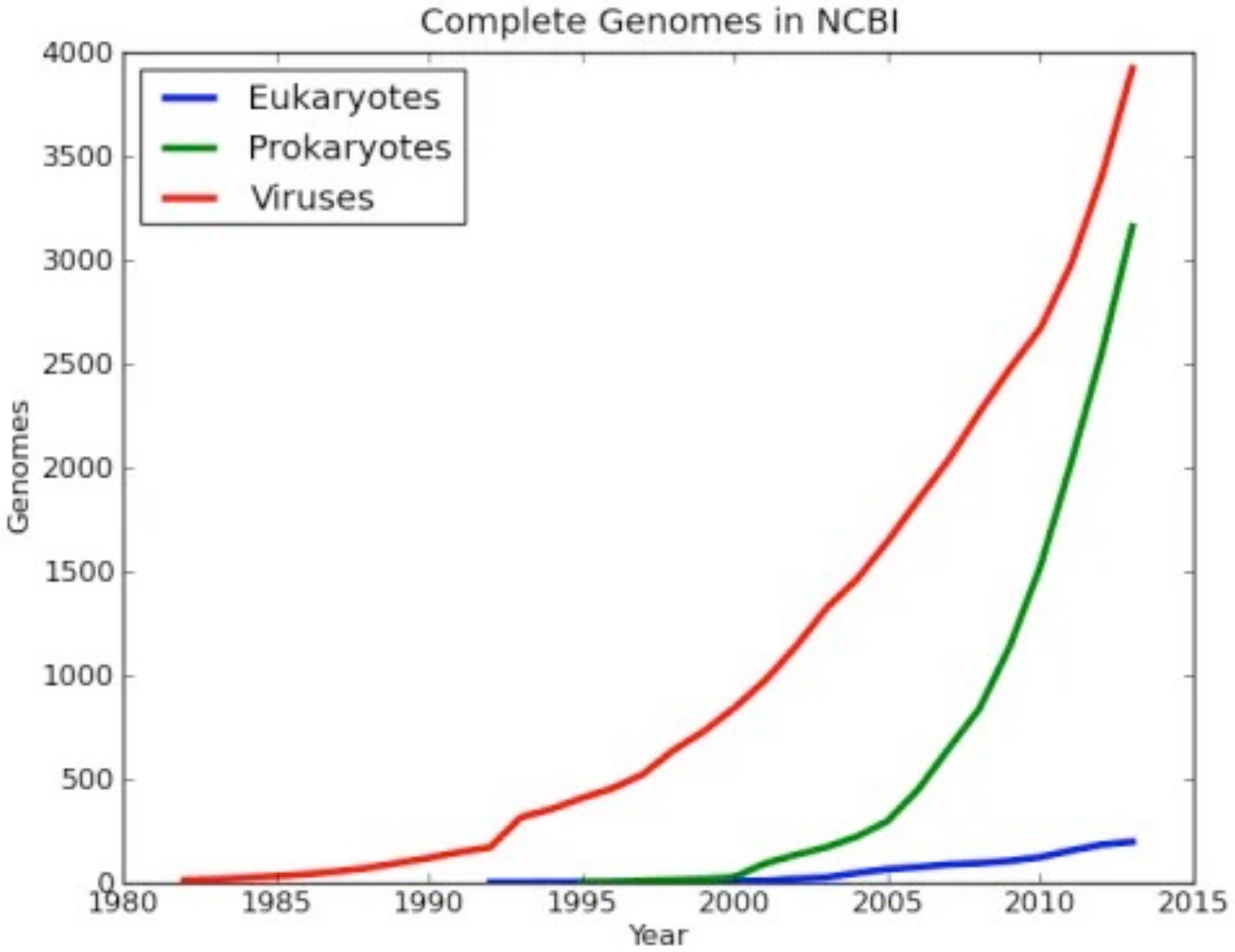
Overview (39406); Eukaryotes (6564); **Prokaryotes (162766)**; Viruses (19743); Plasmids (13734); Organelles (12081)

Filters Download

Choose Columns Page 1 of 3,256 50

#	Organism Name	Organism Groups	Strain	BioSample	BioProjec	Assembly	Leve	Size	
1	'Brassica napus' phytoplasma	Bacteria;Terrabacteria group;Tenericutes	TW1	SAMN09083457	PRJNA464391	GCA_003181115.1		0.743598	2
2	'Candidatus Kapabacteria' thiocyanatum	Bacteria;FCB group;Bacteroidetes/Chloro group	59-99	SAMN05660602	PRJNA279279	GCA_001899175.1		3.27	5
3	'Chrysanthemum coronarium' phytoplasma	Bacteria;Terrabacteria group;Tenericutes	OY-V	SAMD00018609	PRJDB2922	GCA_000744065.1		0.739592	2

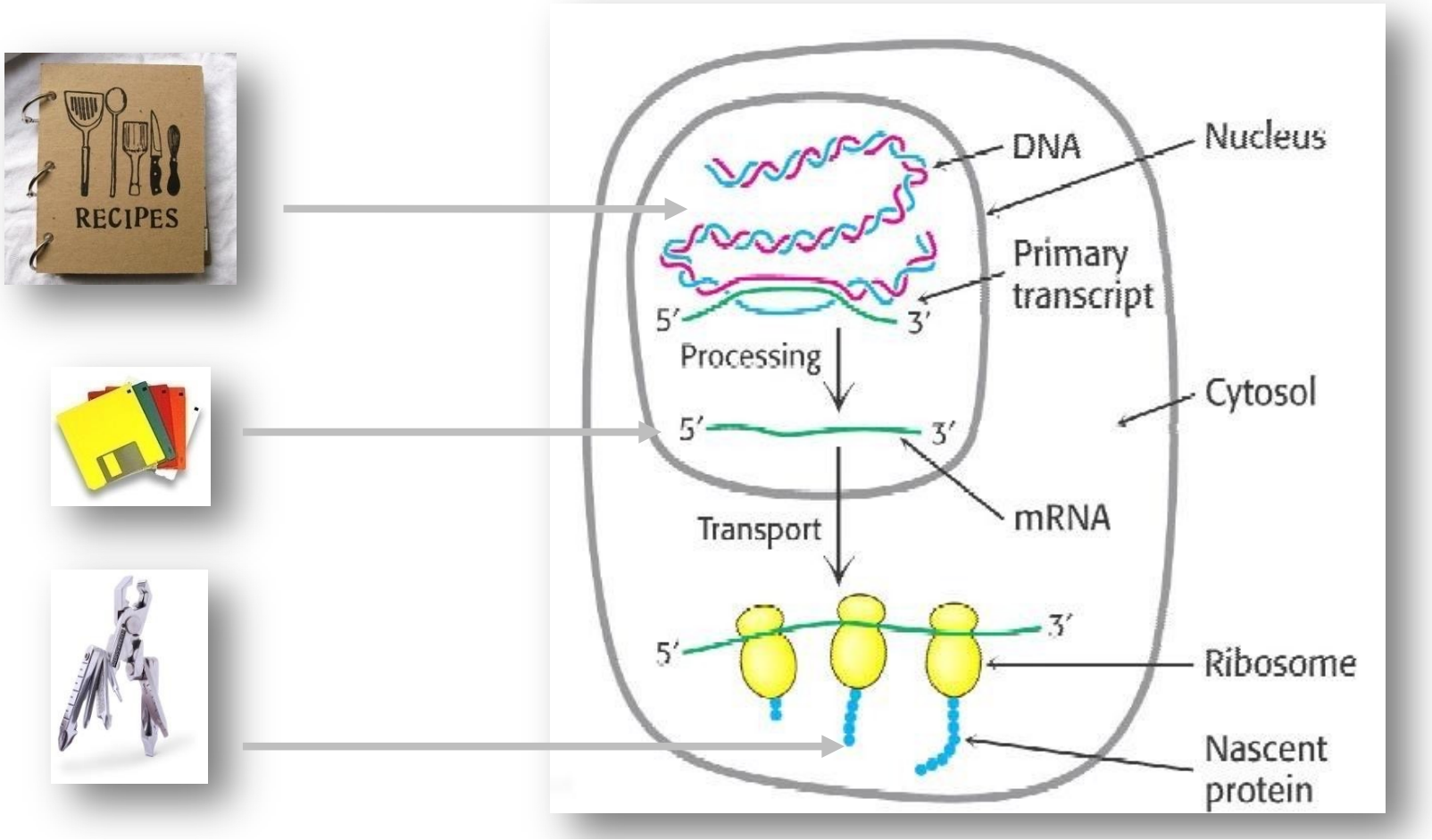
Sequenced genomes at NCBI



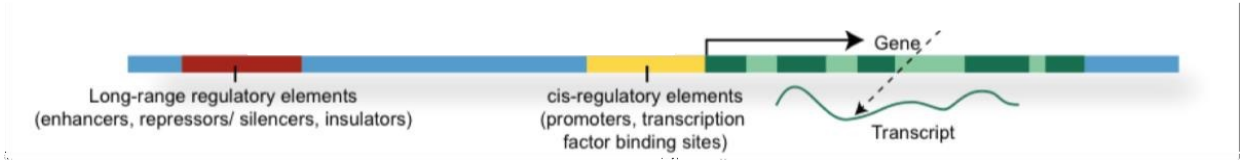
Central dogma

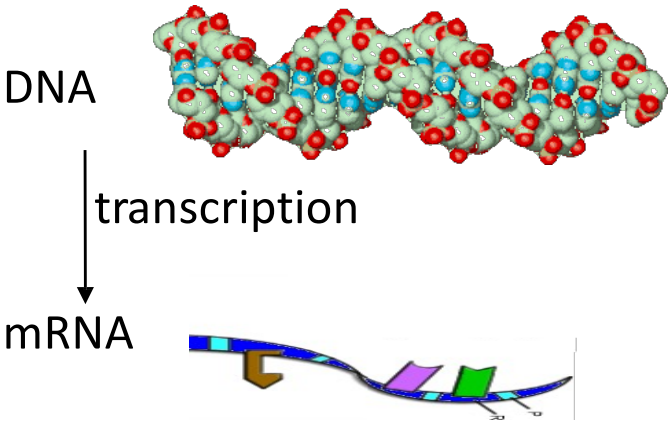
Desde los cromosomas hasta las proteinas…

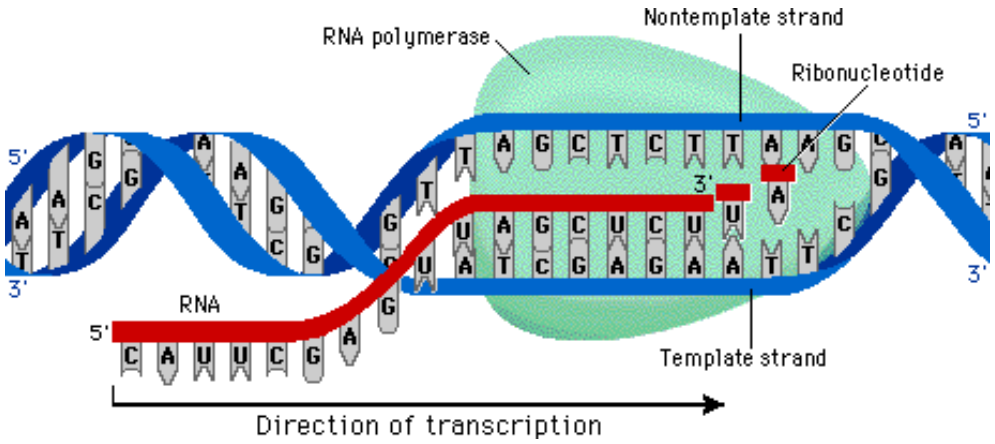
DNA $\xrightarrow{\text{transcription}}$ mRNA $\xrightarrow{\text{translation}}$ protein



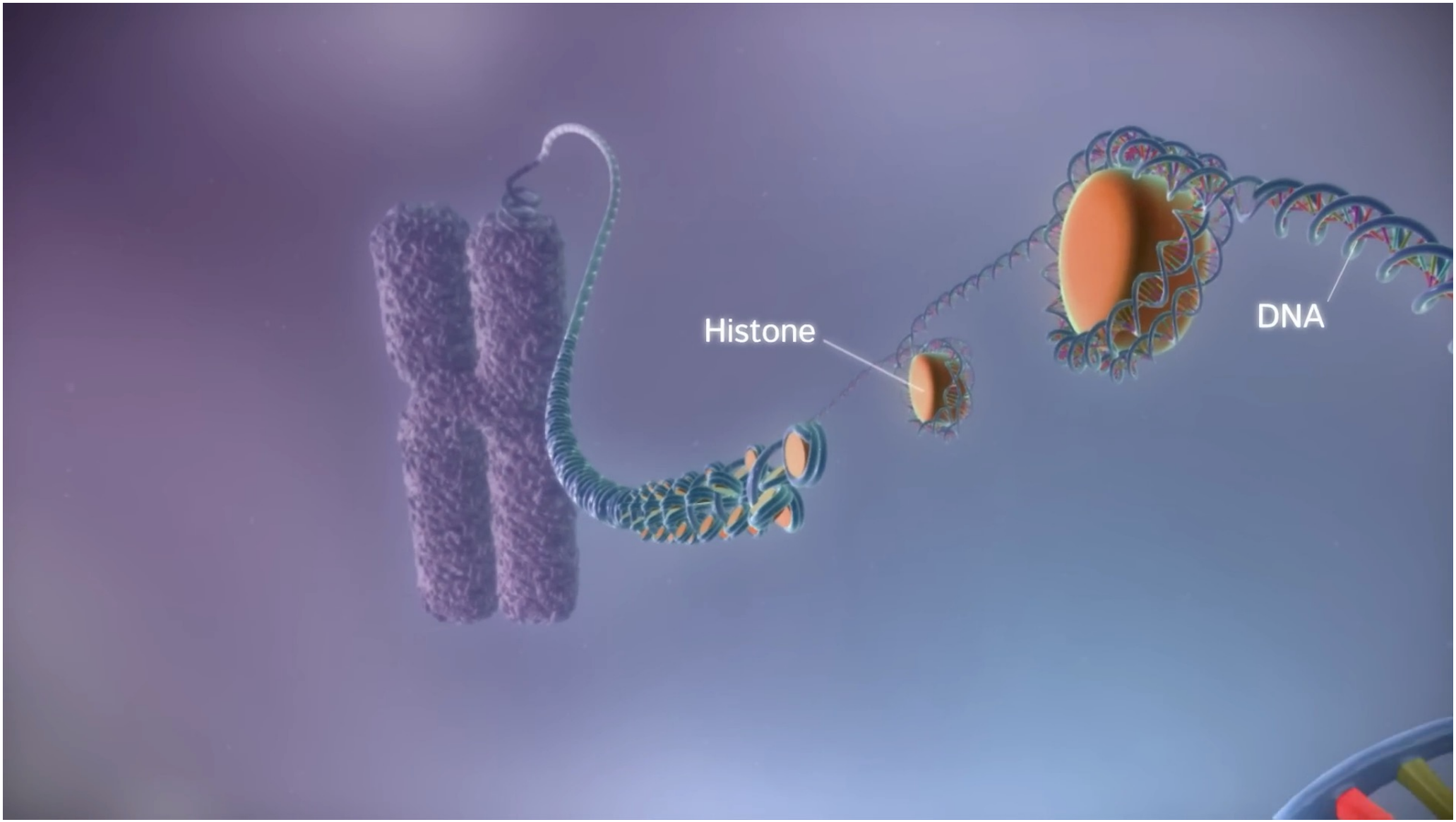
Transcription



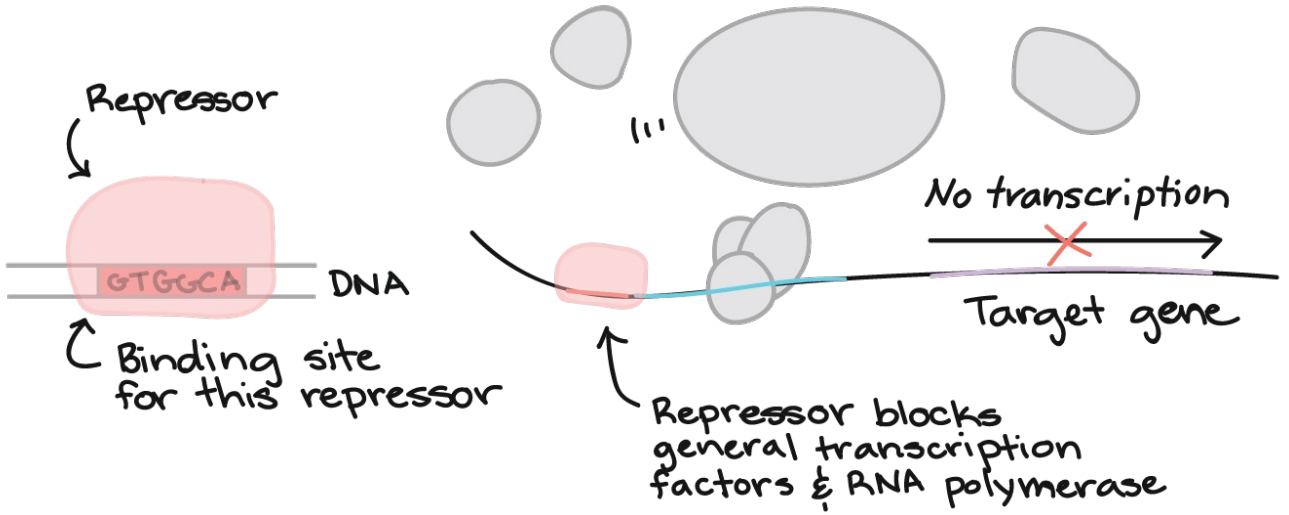


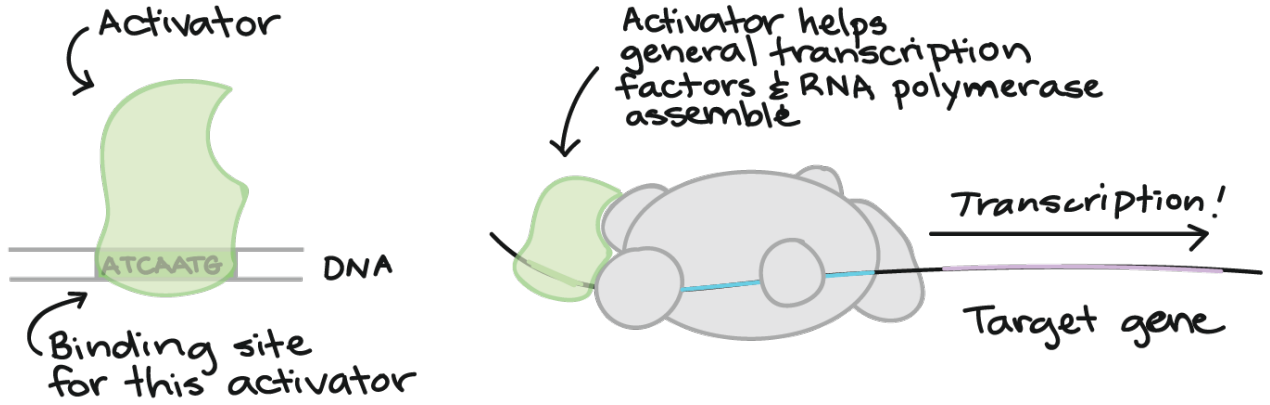


- ARN es similar al ADN pero: nucleotidos que lo componen tienen un azucar adicional, **U**racilo reemplaza a **T**imin, es de hebra única
- Durante la transcripción, la enzima RNA polimerasa recorre el template de ADN y recluta nucleotidos de ARN complementarios.
- Interatúa con proteinas en regiones promotoras (factores de transcripción, reguladores)
- Reconoce sitios de **start** y **stop**
- Finalmete produce una molécula de ARN mensajero: mARN que luego de procesada, es exportada del núcleo.


Histone
DNA

Transcription factors *are proteins that help turn specific genes "on" or "off" by binding to nearby DNA*





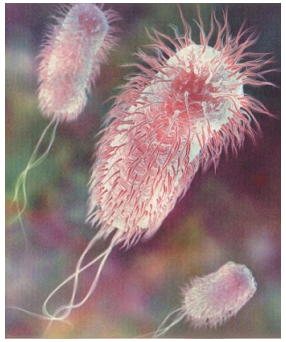
https://www.khanacademy.org

Bacterial models to understand gene regulation

1838–1847 Nucleic Acids Research, 2000, Vol. 28, No. 8

The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12

Ernesto Pérez-Rueda and Julio Collado-Vides*



Ibarra *et al. BMC Genomics* 2013, **14**:126
http://www.biomedcentral.com/1471-2164/14/126

BMC Genomics

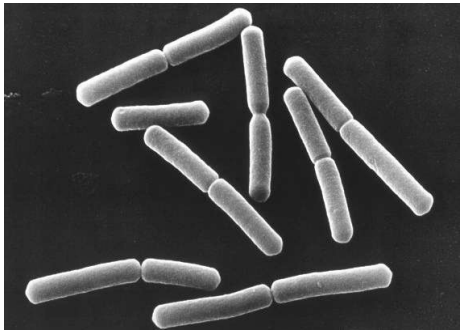
BioMed Central

Research article

Open Access

Identification and analysis of DNA-binding transcription factors in *Bacillus subtilis* and other Firmicutes- a genomic approach

Samadhi Moreno-Campuzano[1], Sarath Chandra Janga[2] and Ernesto Pérez-Rueda*[1]

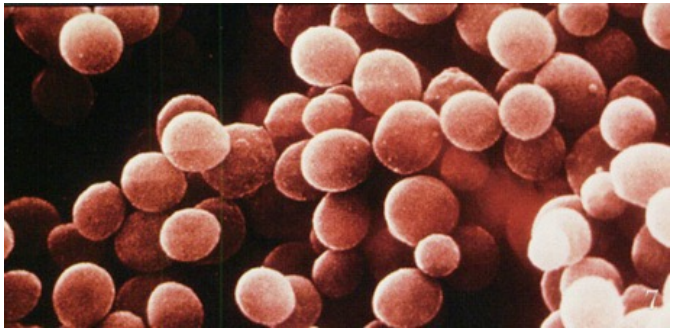


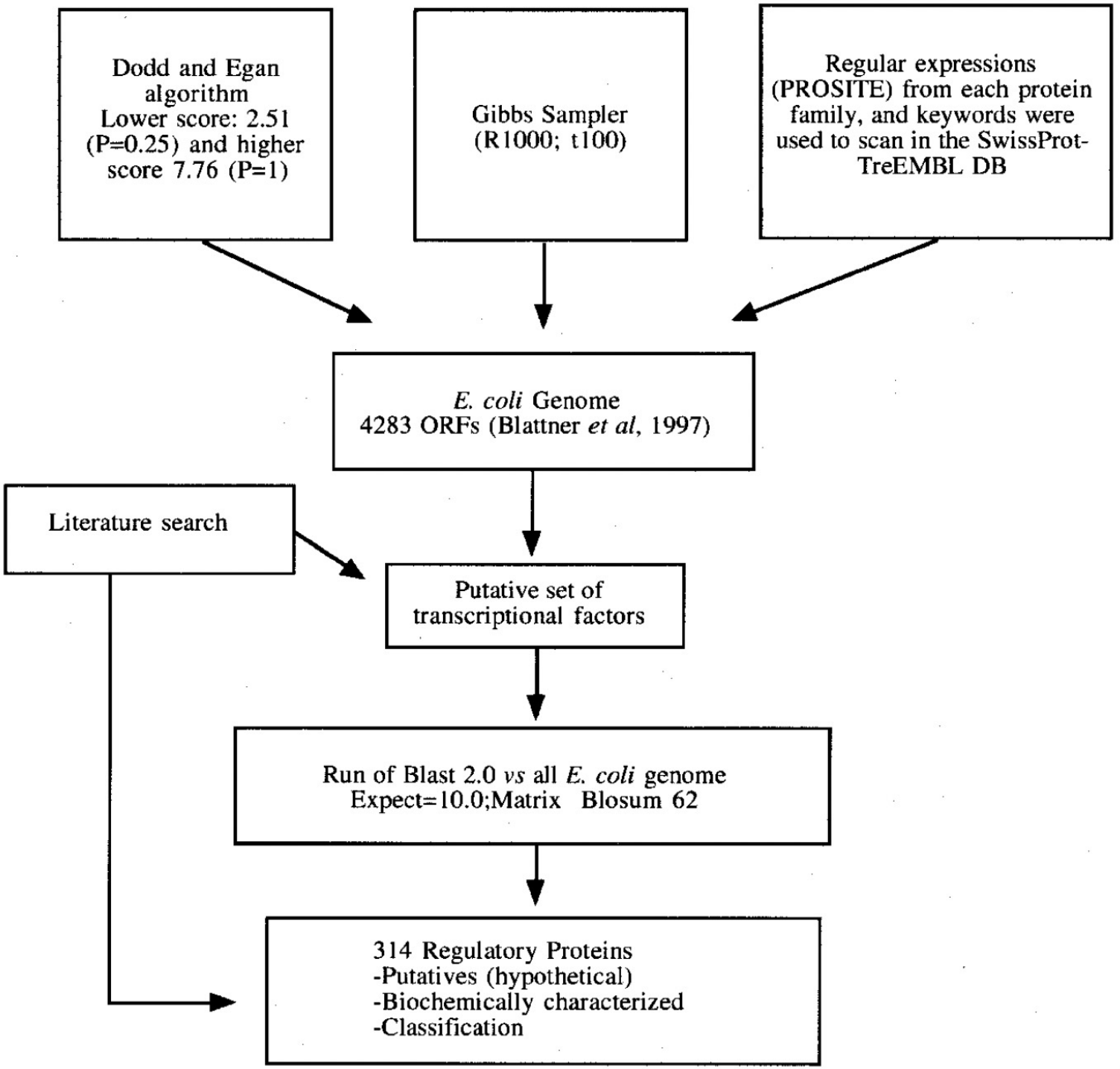
BMC Genomics

RESEARCH ARTICLE Open Access

Global analysis of transcriptional regulators in *Staphylococcus aureus*

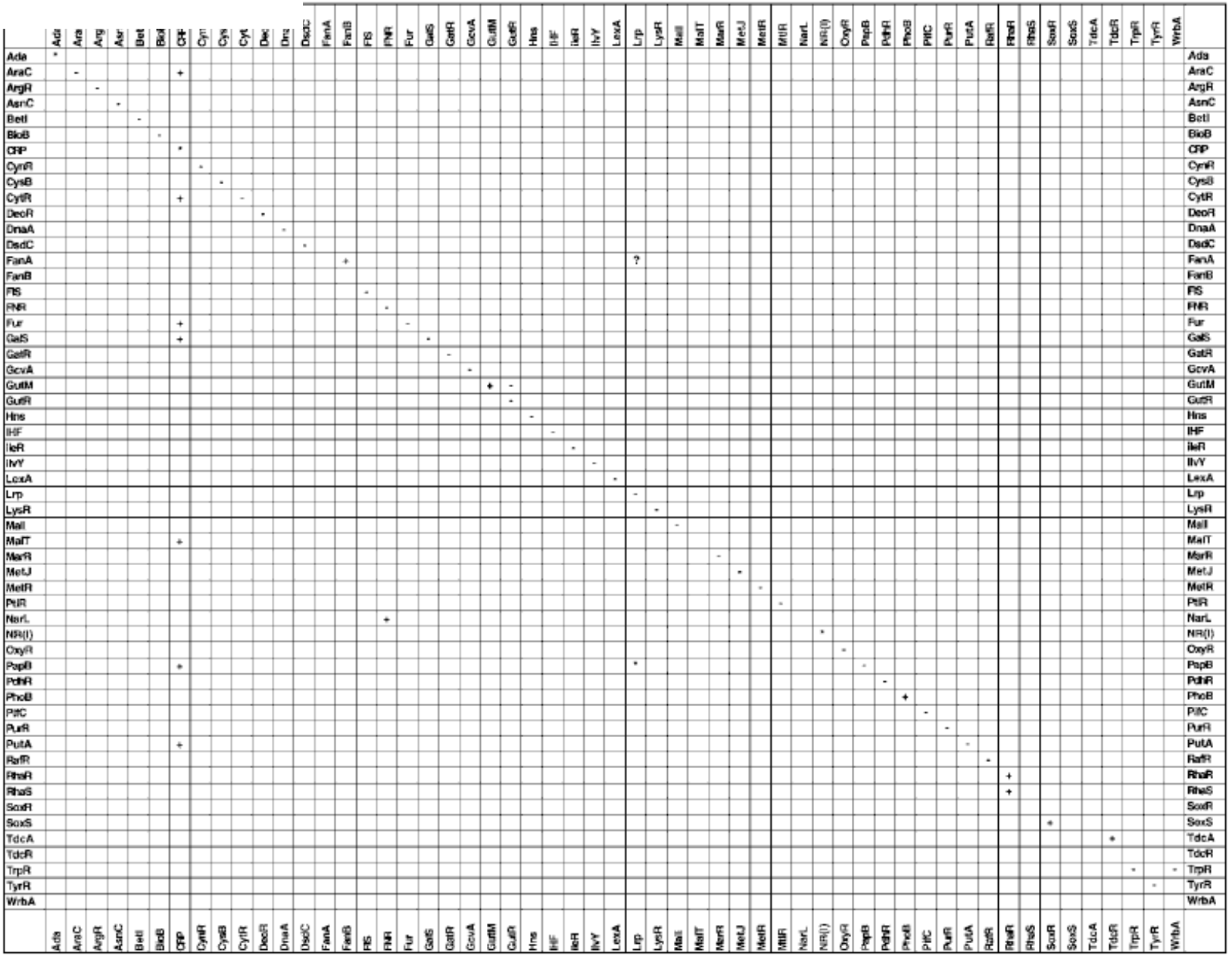
Jose Antonio Ibarra[1,3*], Ernesto Pérez-Rueda[2*], Ronan K Carroll[1] and Lindsey N Shaw[1]




Dodd and Egan algorithm
Lower score: 2.51 (P=0.25) and higher score 7.76 (P=1)
Gibbs Sampler (R1000; t100)
Regular expressions (PROSITE) from each protein family, and keywords were used to scan in the SwissProt-TreEMBL DB
E. coli Genome
4283 ORFs (Blattner et al, 1997)
Literature search
Putative set of transcriptional factors
Run of Blast 2.0 vs all E. coli genome
Expect=10.0;Matrix Blosum 62
314 Regulatory Proteins
-Putatives (hypothetical)
-Biochemically characterized
-Classification

From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*

Denis Thieffry,* Araceli M. Huerta, Ernesto Pérez-Rueda, and Julio Collado-Vides

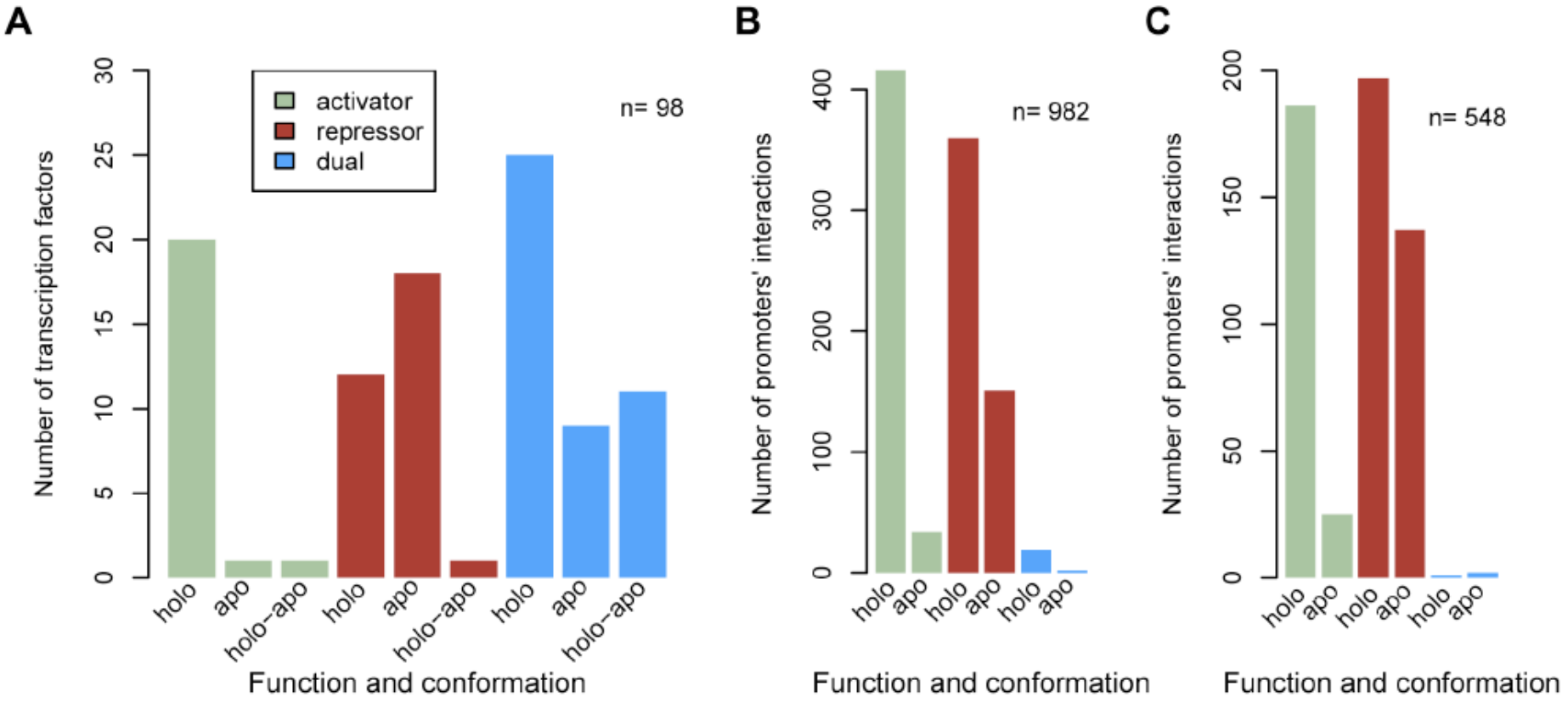


OPEN ACCESS Freely available online

PLOS ONE

Transcription Factors in *Escherichia coli* Prefer the *Holo* Conformation

Yalbi Itzel Balderas-Martínez[1]*, Michael Savageau[2], Heladia Salgado[1], Ernesto Pérez-Rueda[3], Enrique Morett[3], Julio Collado-Vides[1]*



RESEARCH ARTICLE Open Access

Lessons from the modular organization of the transcriptional regulatory network of Bacillus subtilis

Julio A Freyre-González[1†], Alejandra M Manjarrez-Casas[2†], Enrique Merino[2], Mario Martinez-Nuñez[2], Ernesto Perez-Rueda[3] and Rosa-María Gutiérrez-Ríos[2*]

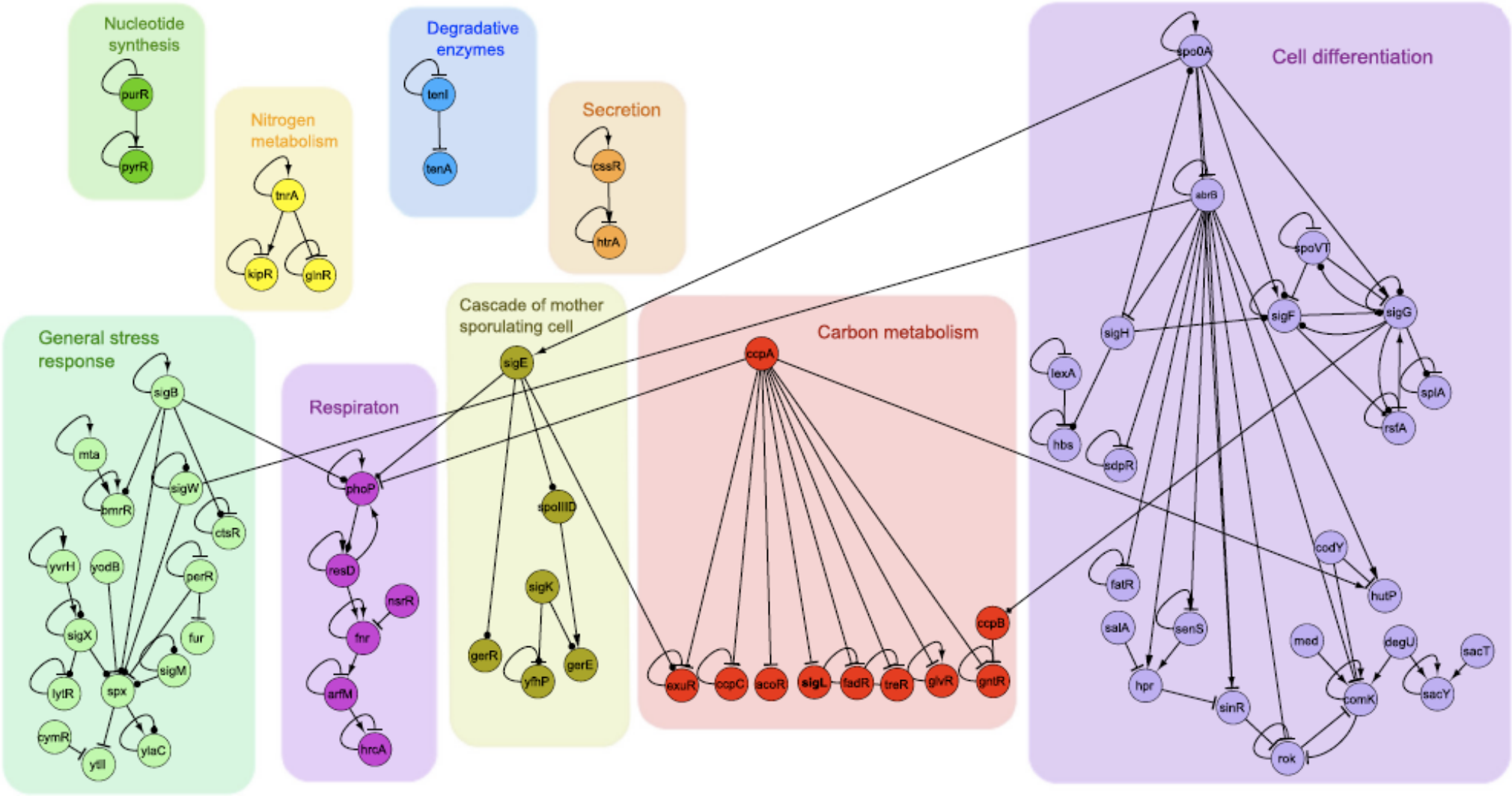
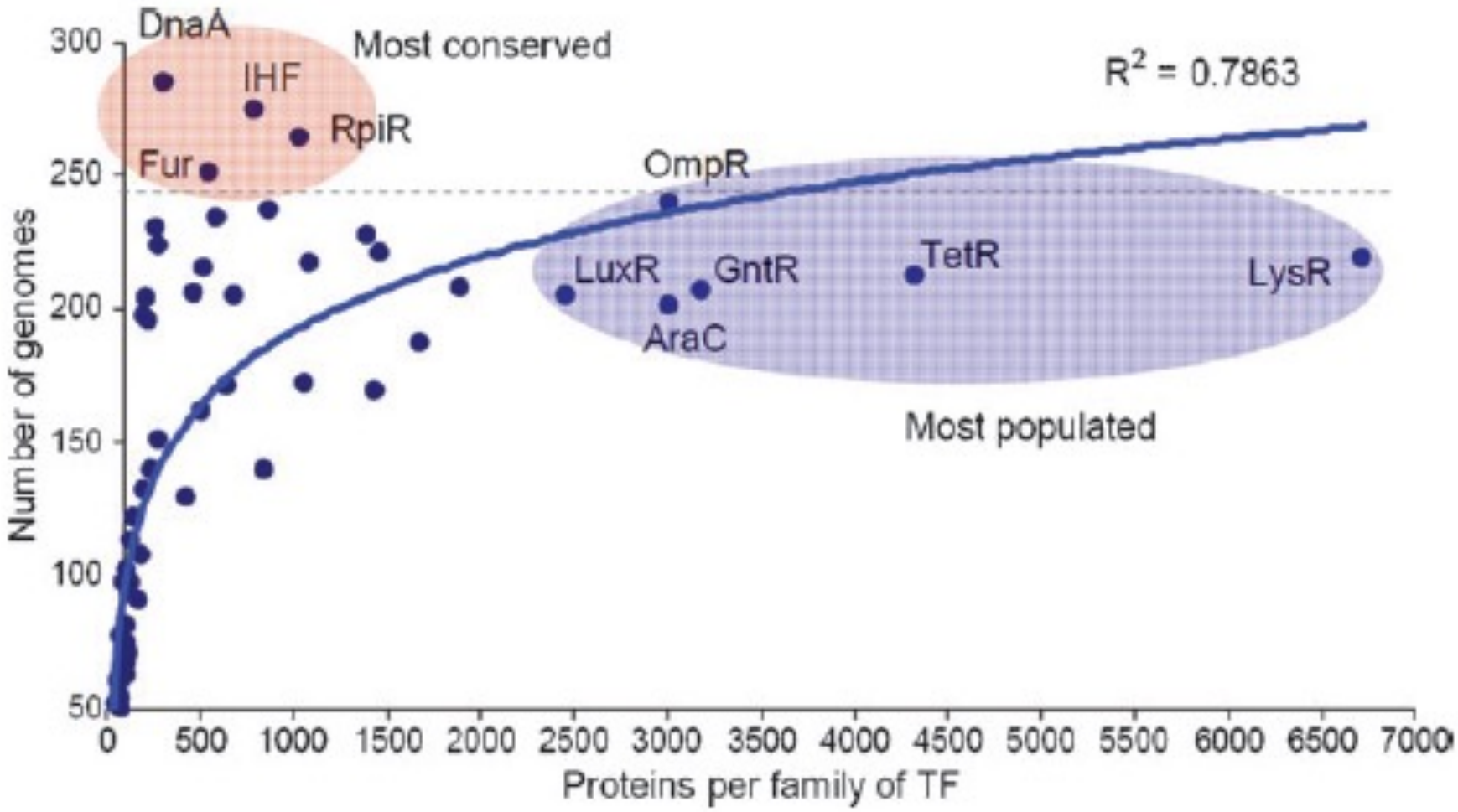


Figure 2 Cross-talk between modules. Master regulators (hubs) interconnect functional modules. At the higher levels, each master regulator is indicated. The color of each TF relates it to the module to which it belongs. We also show (top left) four disconnected groups that represent modules that are not interconnected by master regulators.

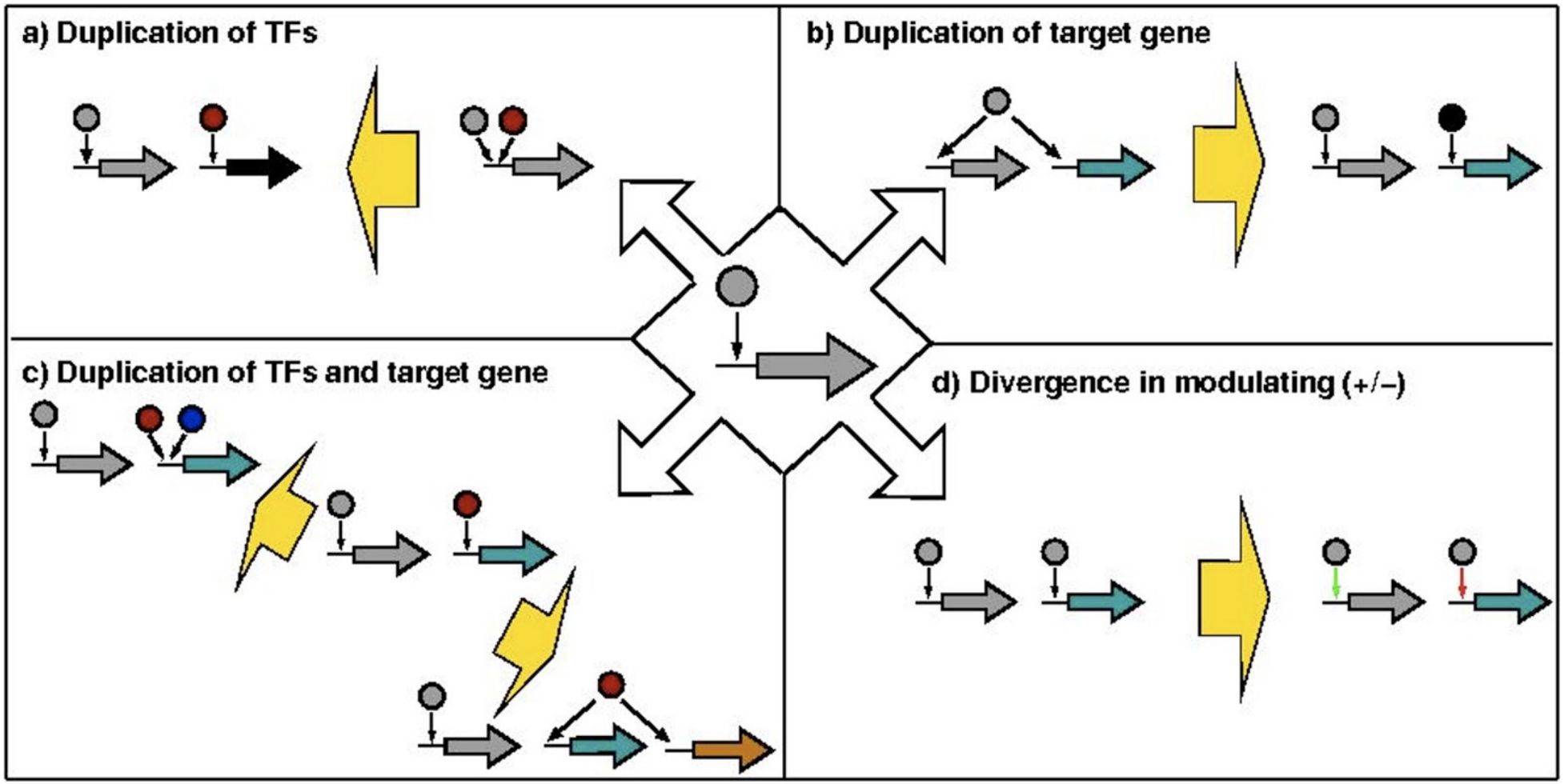
Scaling relationship in the gene content of transcriptional machinery in bacteria†‡

Ernesto Pérez-Rueda,[a] Sarath Chandra Janga*[b] and Agustino Martínez-Antonio*[c]



New insights into the regulatory networks of paralogous genes in bacteria

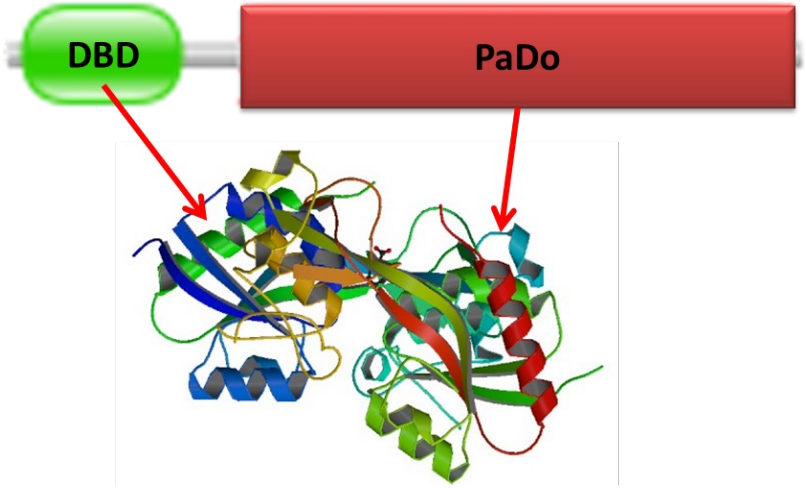
Mario A. Martínez-Núñez,[1] Ernesto Pérez-Rueda,[2]
Rosa María Gutiérrez-Ríos[1] and Enrique Merino[1]

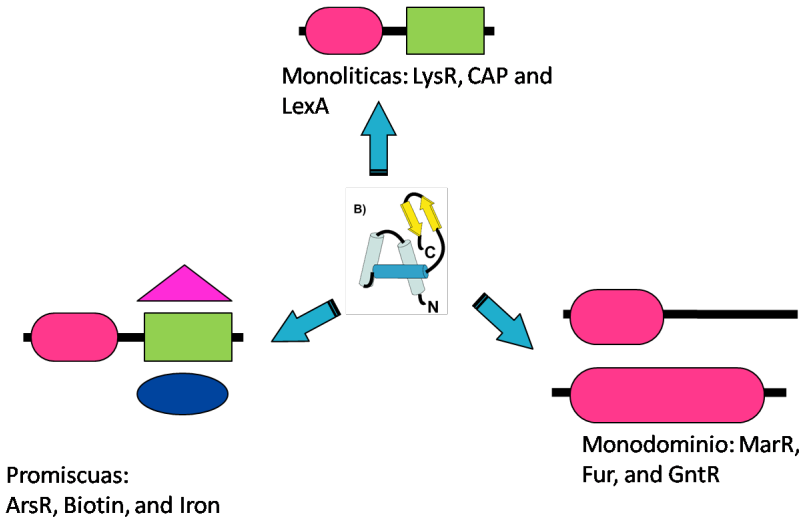


Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria

Nancy Rivera-Gómez, Lorenzo Segovia and Ernesto Pérez-Rueda

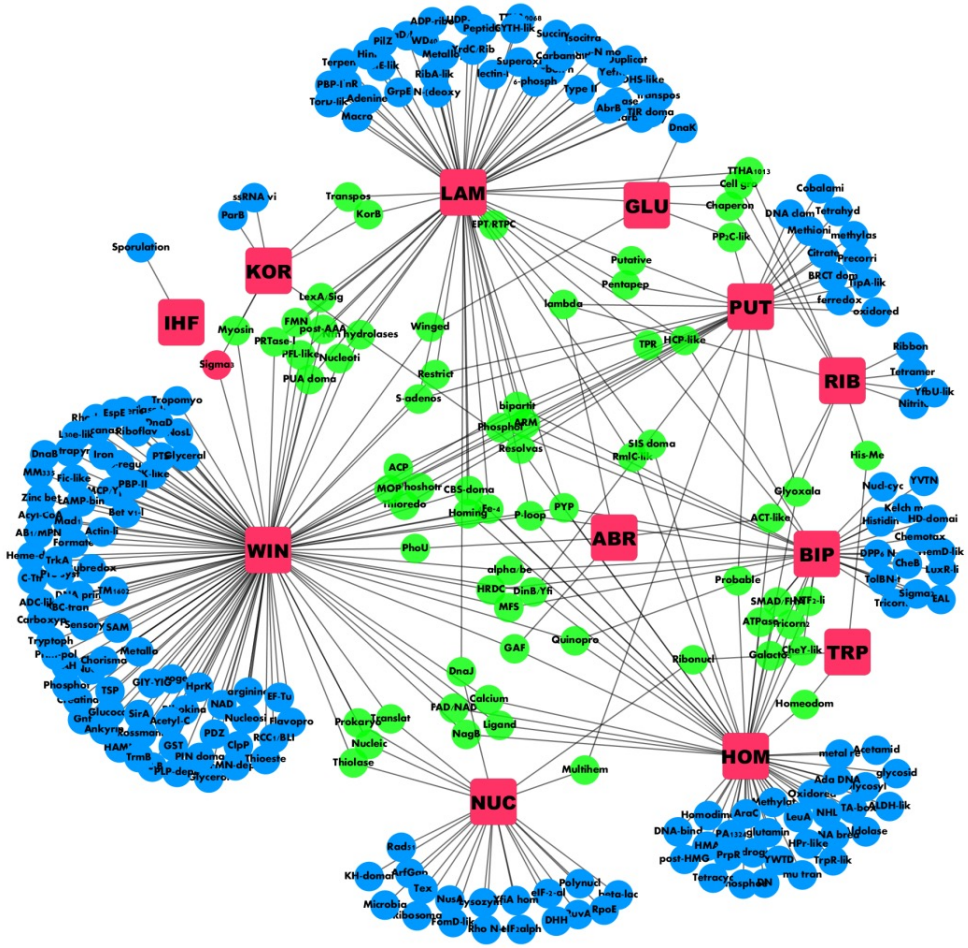
TFs have (in general) two domins: DBD (DNA-binding domain) & a Partner Domain (PaDo)





In bacteria and archaea, 12 superfamiles of DBDs have been identified, associated with 100 families

Domain organization (*shuffling*)



DBD

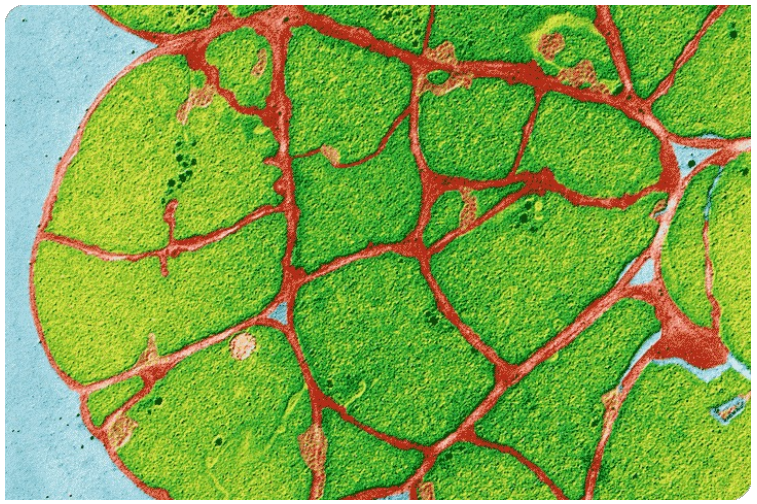
PaDos exclusive

PaDos share with…

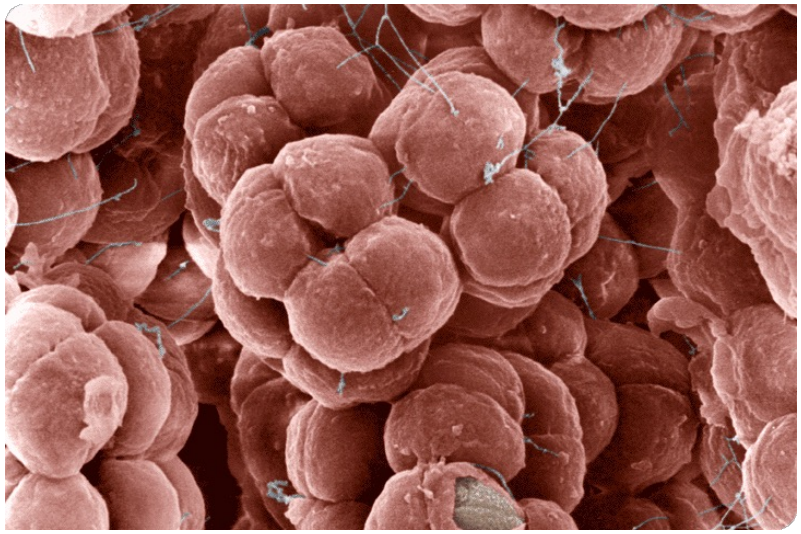
- The DBD is highly conserved along bacteria and archaea.
- PaDos are associated with the DBDs at different proportions

Gene regulation in Archaea

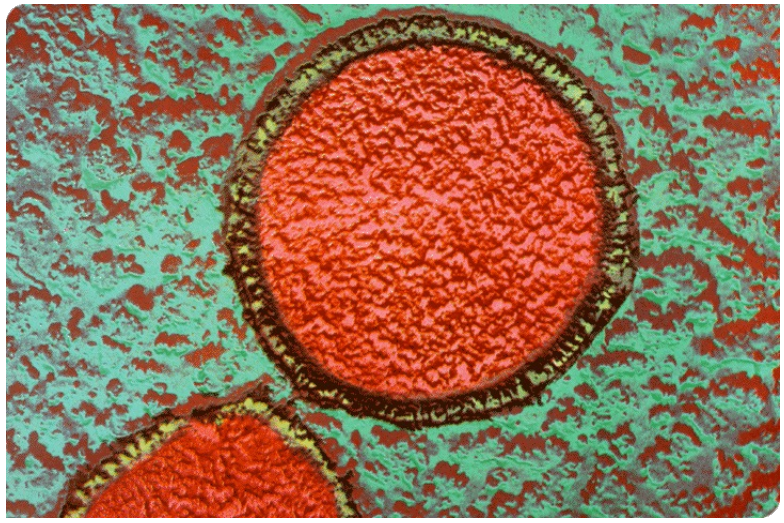
Sulfolobus solfataricus



Methanosarcina rumen



Halococcus salifodinae



Staphylothermus marinus

Methanogenic *Archaea* and human periodontal disease

Paul W. Lepp[†‡§¶], Mary M. Brinig[†§], Cleber C. Ouverney[†‡§], Katherine Palm[‡], Gary C. Armitage[‖], and David A. Relman[†‡§]

Departments of [†]Microbiology and Immunology and [‡]Medicine, Stanford University, Stanford, CA 94305; [§]Veterans Affairs Palo Alto Health Care System, Palo Alto, CA 94304; and [‖]Department of Stomatology, University of California, San Francisco, CA 94143

Edited by Stanley Falkow, Stanford University, Stanford, CA, and approved February 23, 2004 (received for review December 31, 2003)

Arch Microbiol. 1982 Feb;131(1):14-8.

Enumeration of Methanobrevibacter smithii in human feces.

Miller TL, Wolin MJ.

J Clin Microbiol. 1990 July; 28(7): 1666-1668

Methanogenic bacteria in human vaginal samples.

N Belay, B Mukhopadhyay, E Conway de Macario, R Galask and L Daniels

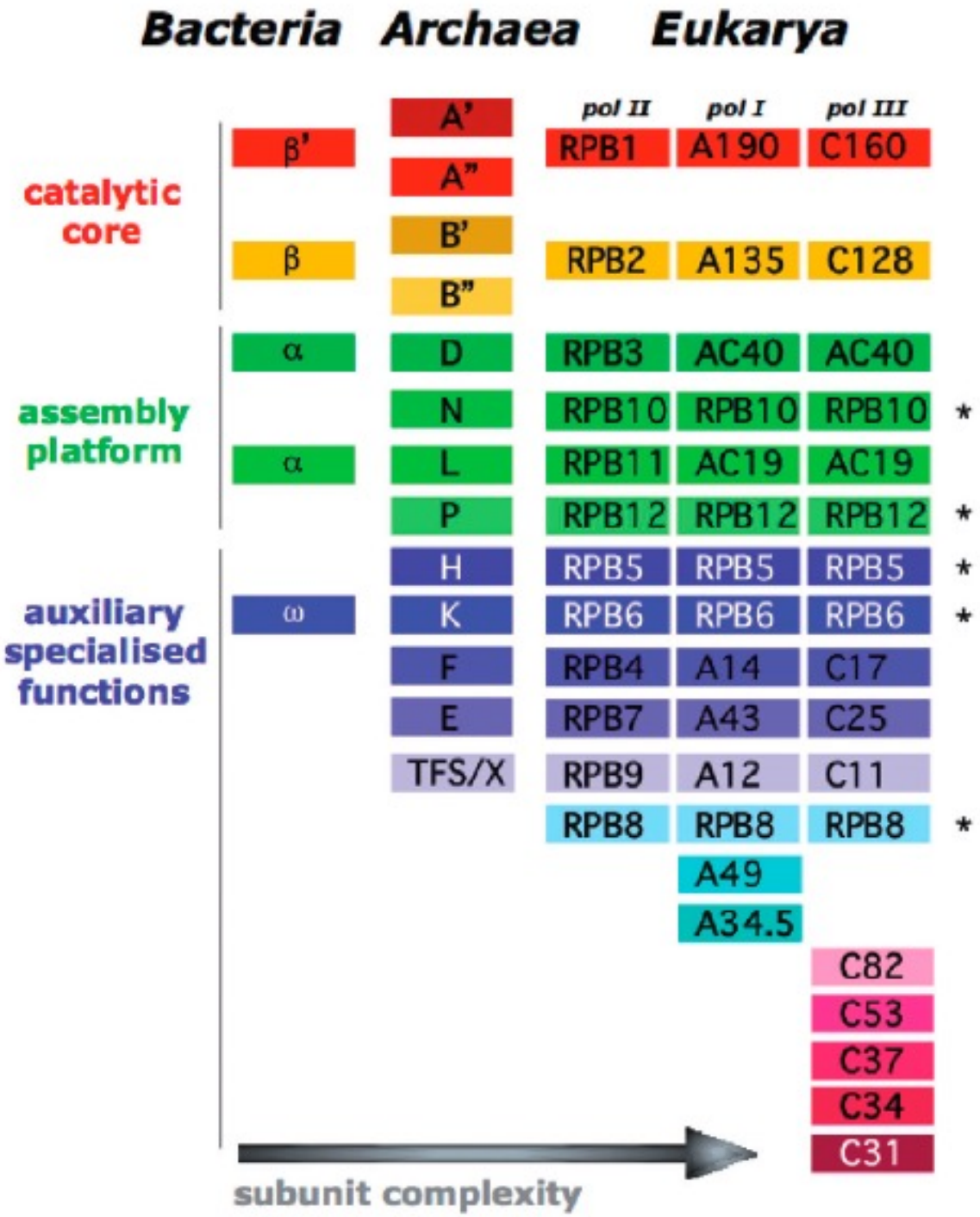
OPEN ACCESS Freely available online

PLOS ONE

Archaea on Human Skin

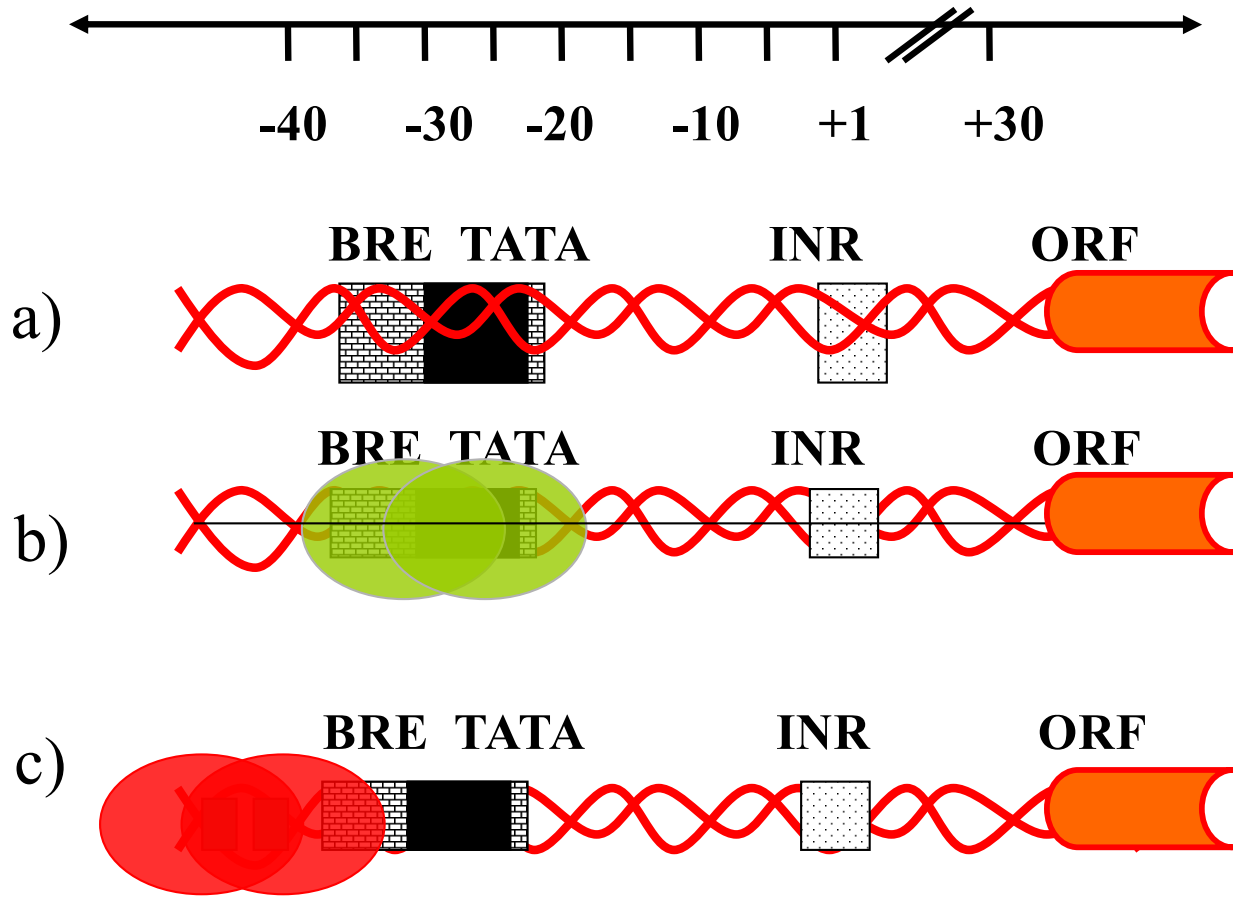
Alexander J. Probst, Anna K. Auerbach, Christine Moissl-Eichinger*

Institute for Microbiology and Archaea Center, University of Regensburg, Regensburg, Germany

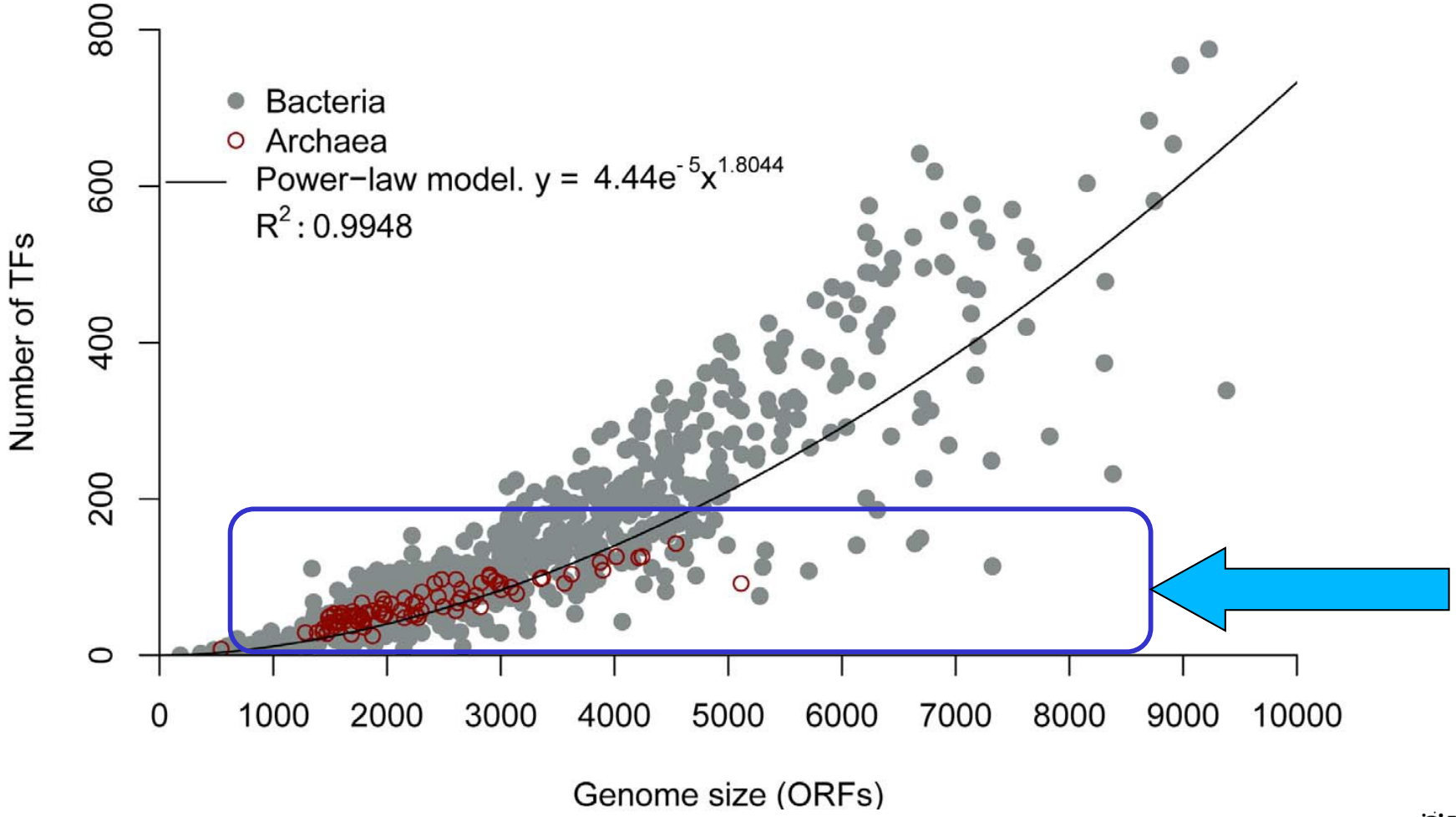


* Conserved in archaea and eukarya

Regulatory structure



The number of TFs increase as a function of genome size



OPEN ACCESS Freely available online

PLOS ONE

Increments and Duplication Events of Enzymes and Transcription Factors Influence Metabolic and Regulatory Diversity in Prokaryotes

Mario Alberto Martínez-Núñez[1]*, Augusto Cesar Poot-Hernandez[2], Katya Rodríguez-Vázquez[1], Ernesto Perez-Rueda[2]*

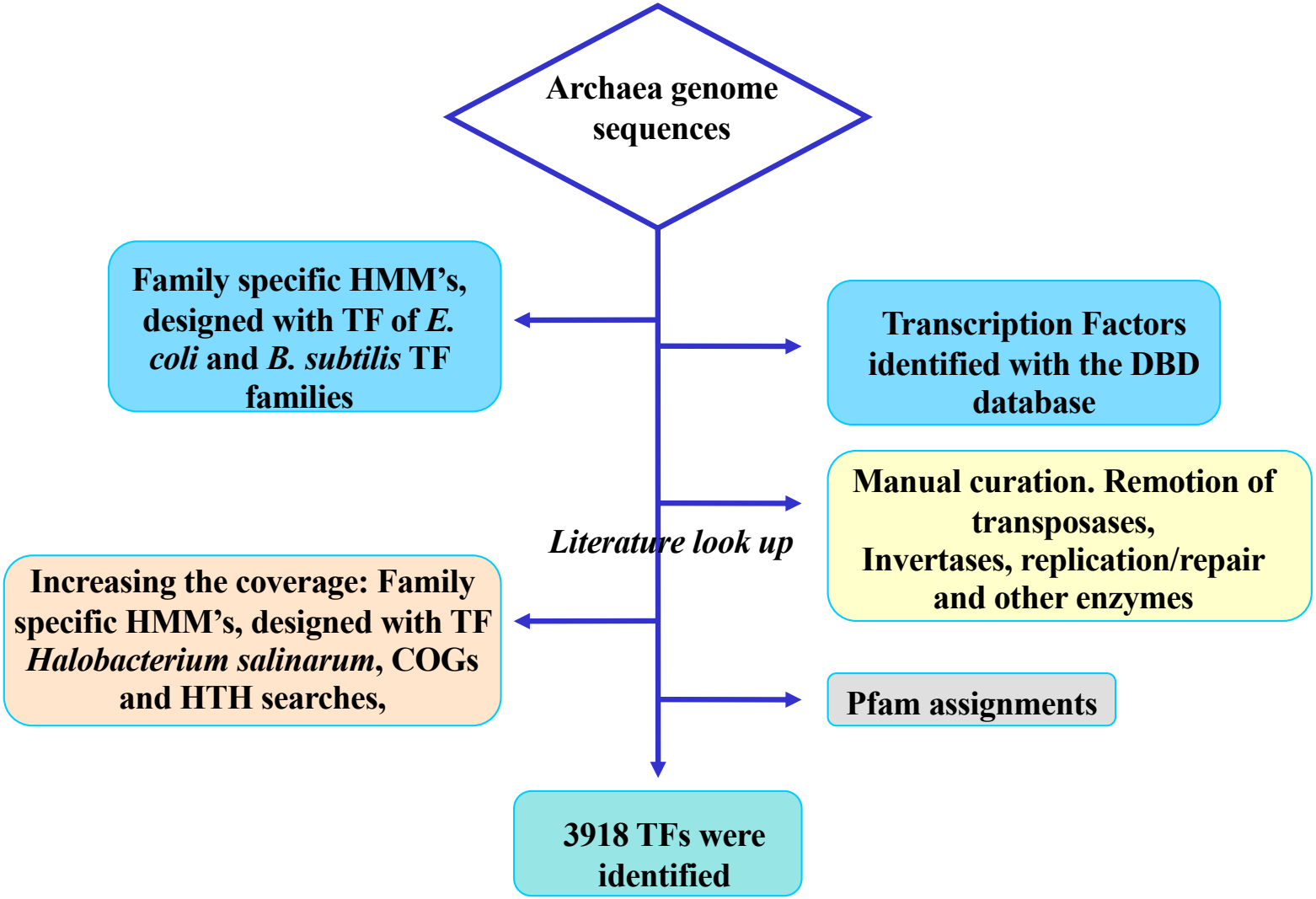
Questions…

¿What is the diversity of TFs?

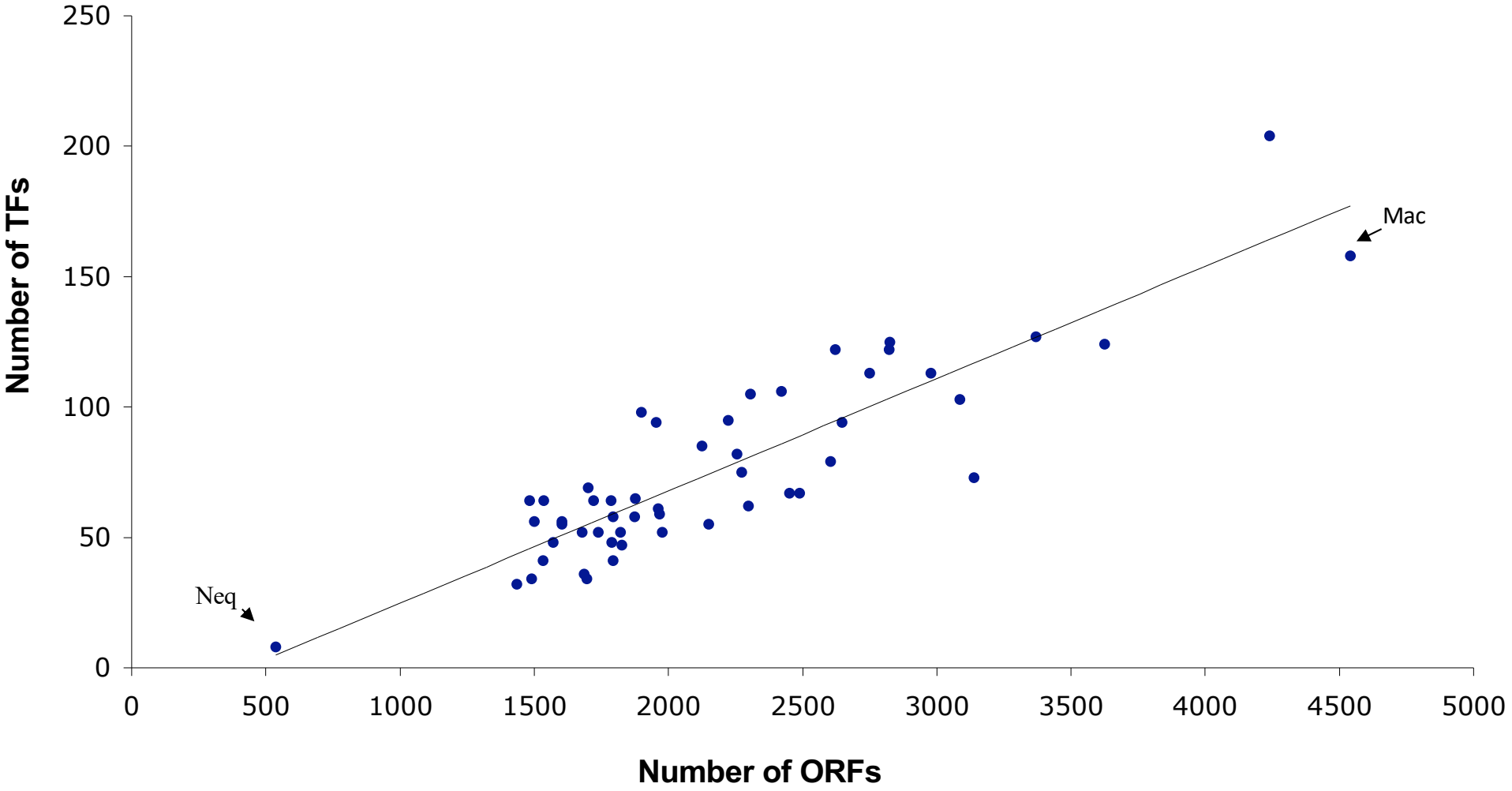
¿How do they regulate the gene expression?

¿What is their regulatory role?

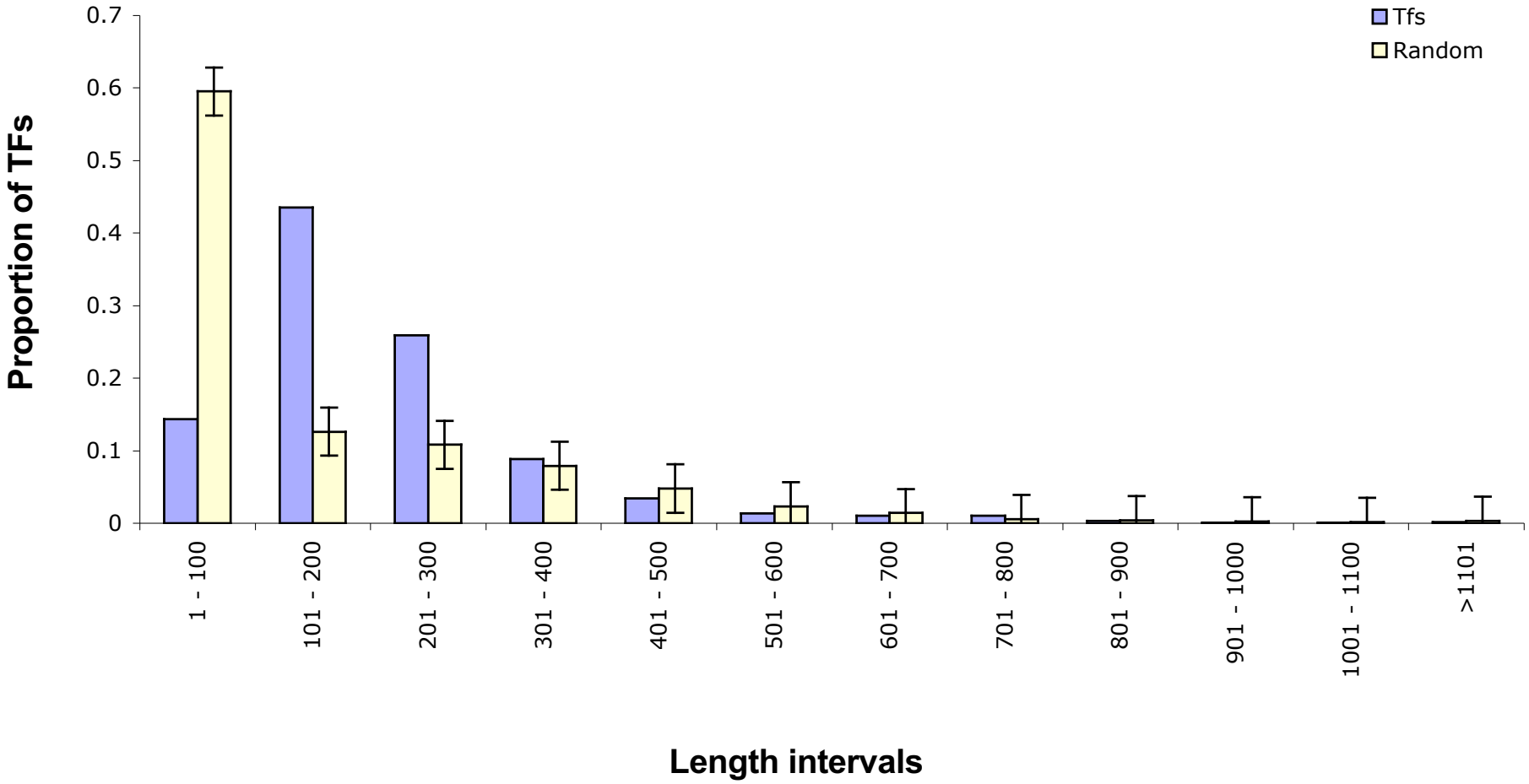
Identification of TFs



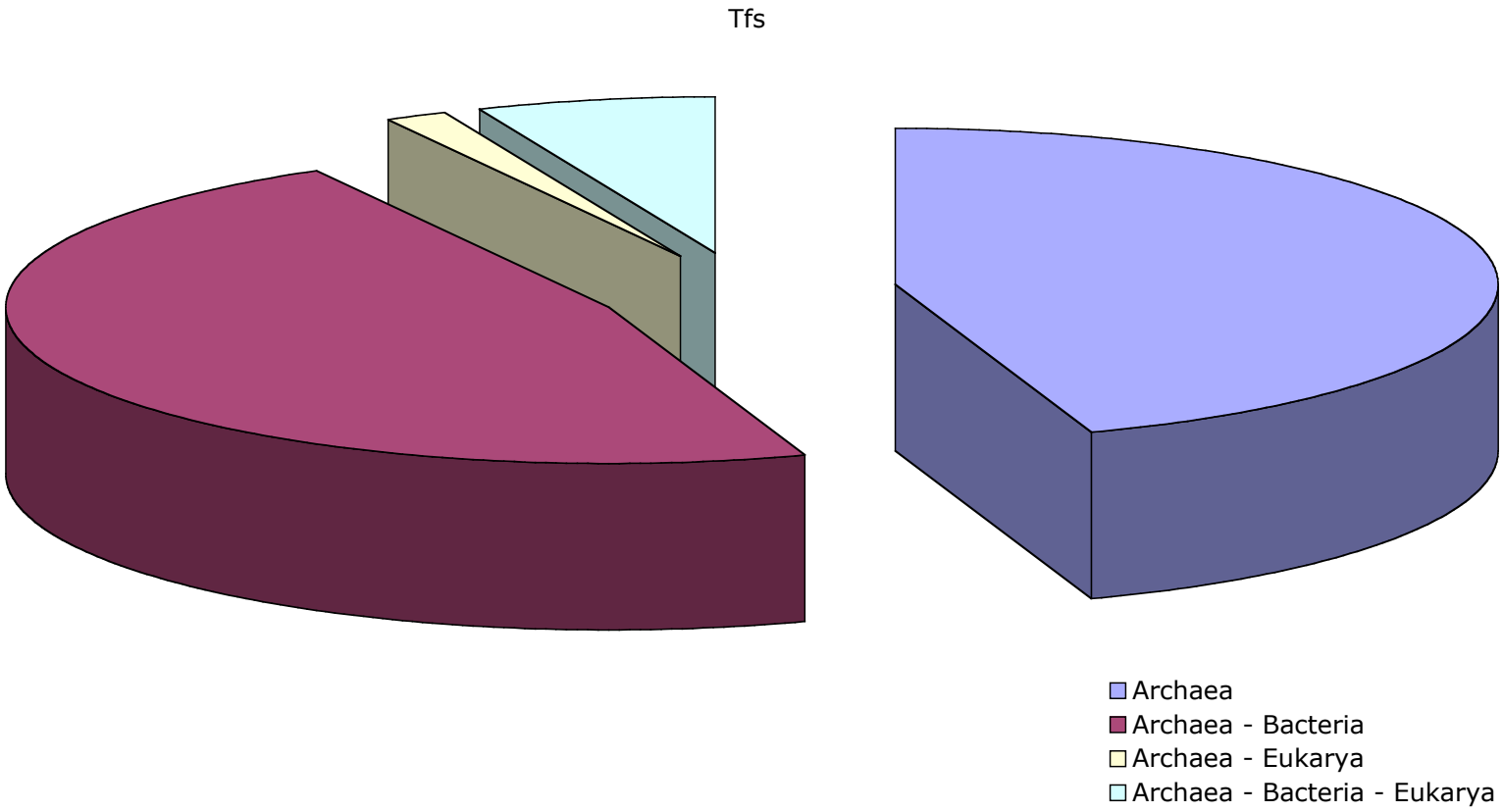
Abundance of TFs in archaea



Lenght sizes…



Archaeal TFs are homologous to bacterial and eukaryal TFs



Hypothesis

Archaeal TFs form multiple complex as eukaryal transcription factors, although they are similar at sequence level to bacterial ones

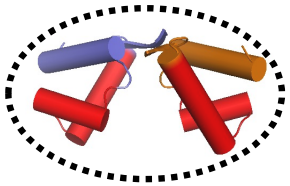
Identification and Genomic Analysis of Transcription Factors in Archaeal Genomes Exemplifies Their Functional Architecture and Evolutionary Origin

Ernesto Pérez-Rueda*,[1] and Sarath Chandra Janga*,[2]

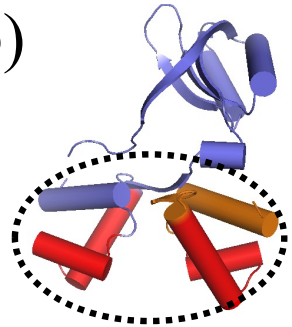
[1]Departamento de Ingeniería Celular y Biocatálisis, IBT-UNAM, AP 565-A, Cuernavaca, Morelos, México

[2]MRC Laboratory of Molecular Biology, Cambridge, United Kingdom

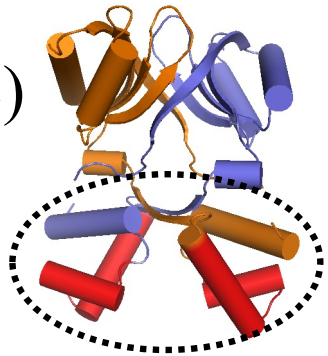
a)



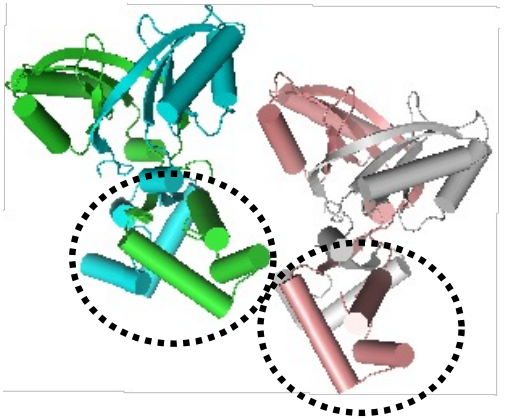
b)



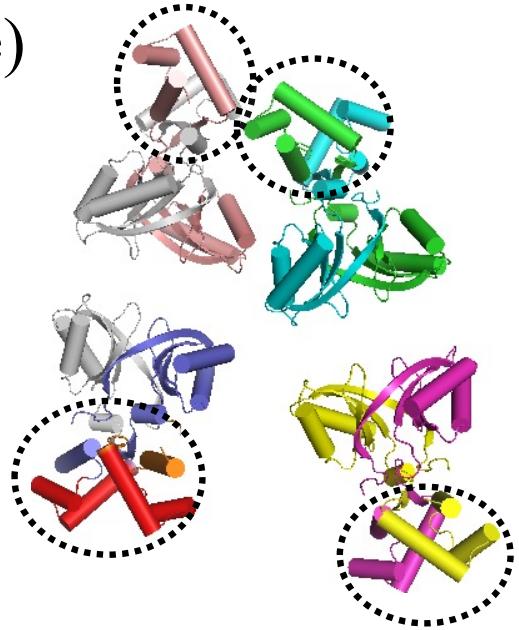
c)



d)



e)



A) DBD => Homodimer - feast/famine regulatory proteins(FFRPs).

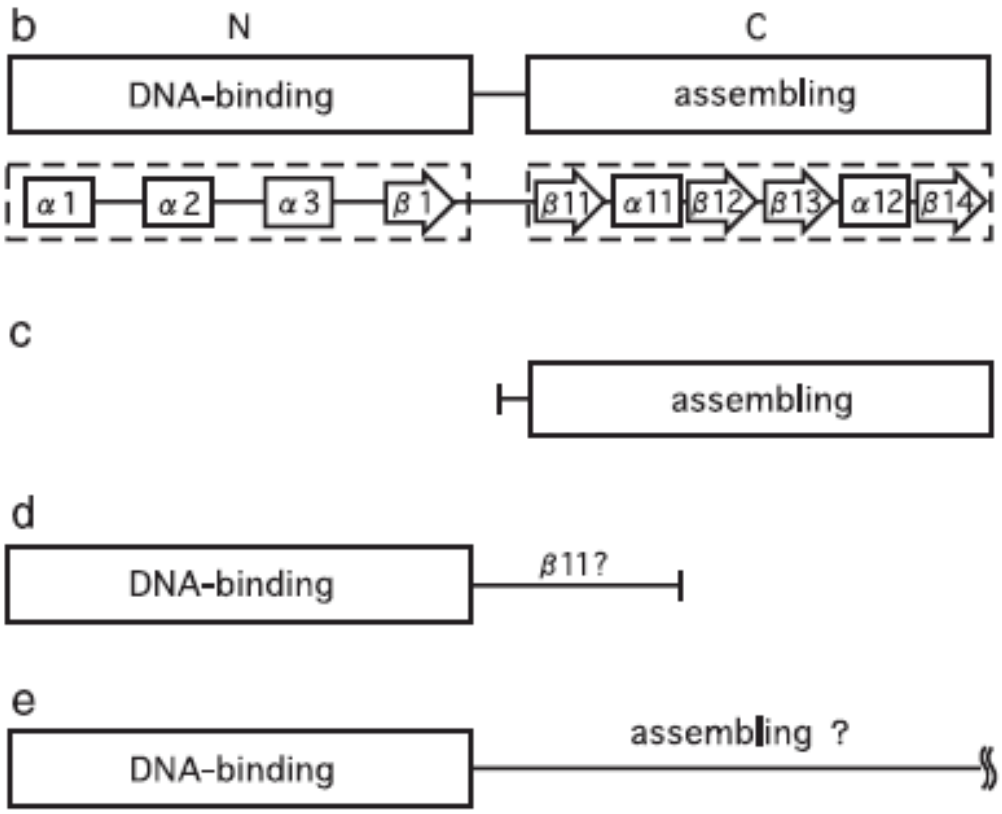
B) DBD - Partner Domain.

C) Two domain protein forms homodimer

D) heterotetramer

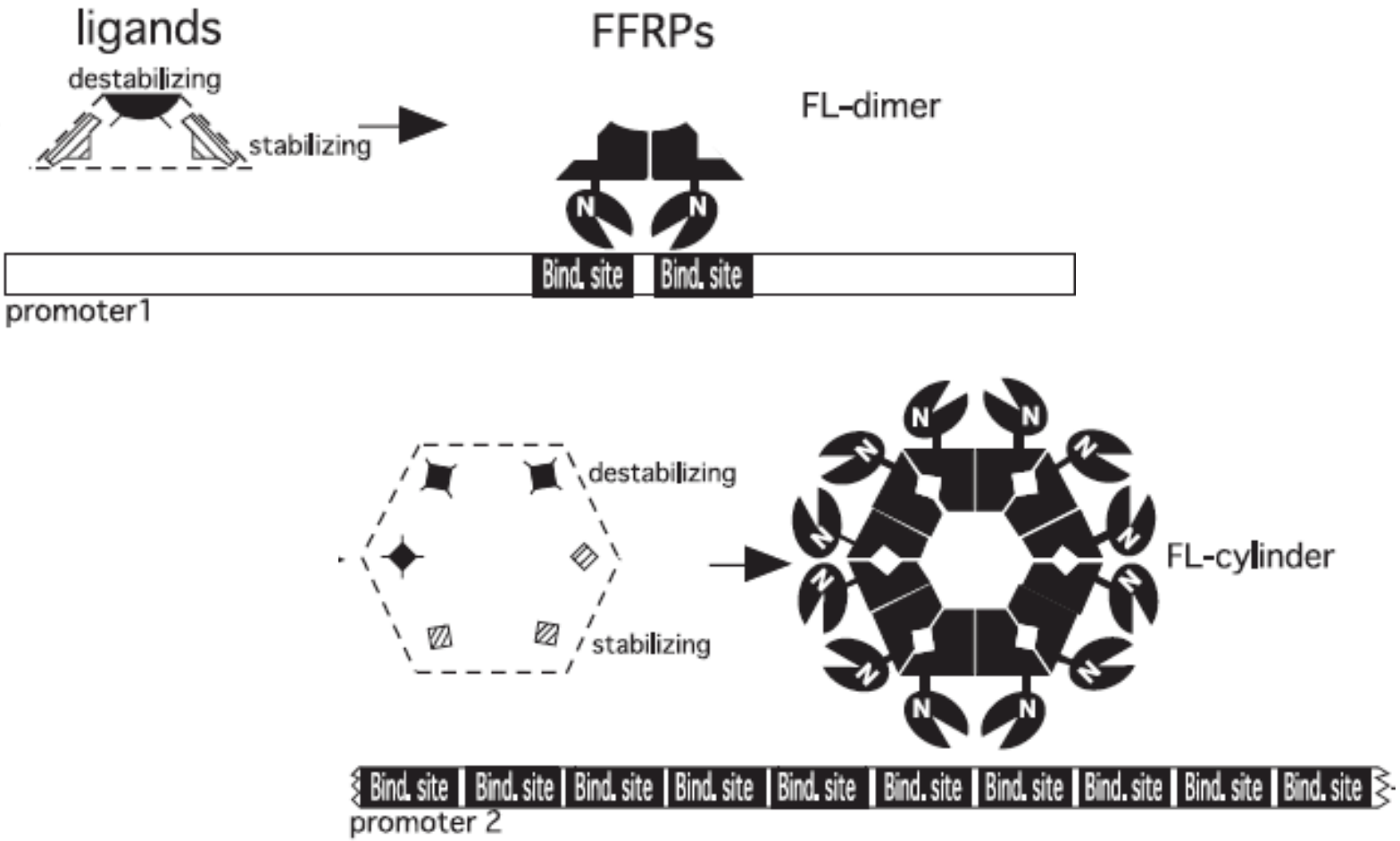
E) Octamer

Archaeal TFs

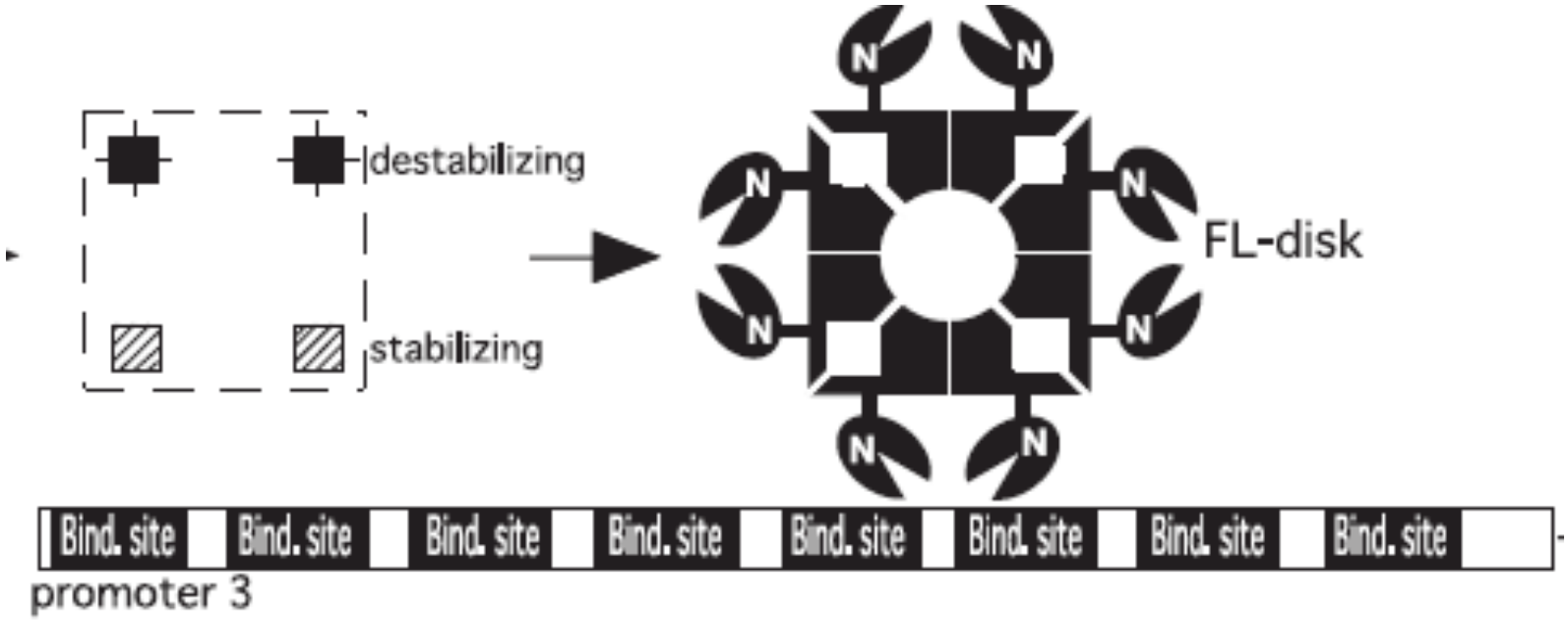


FL11 in *Pyrococcus* sp forms dodecamers, tetramers or they work as histone-like proteins

Multiple complexes



Multiple complexes



TrmB de *Pyrococcus furiosus*, es un tetramero a temperatura ambiente y un octámero con su inductor (maltotriosa o maltosa)

TrmB binds to multiple sugars (maltose, sacarose, maltotriose & trealose)

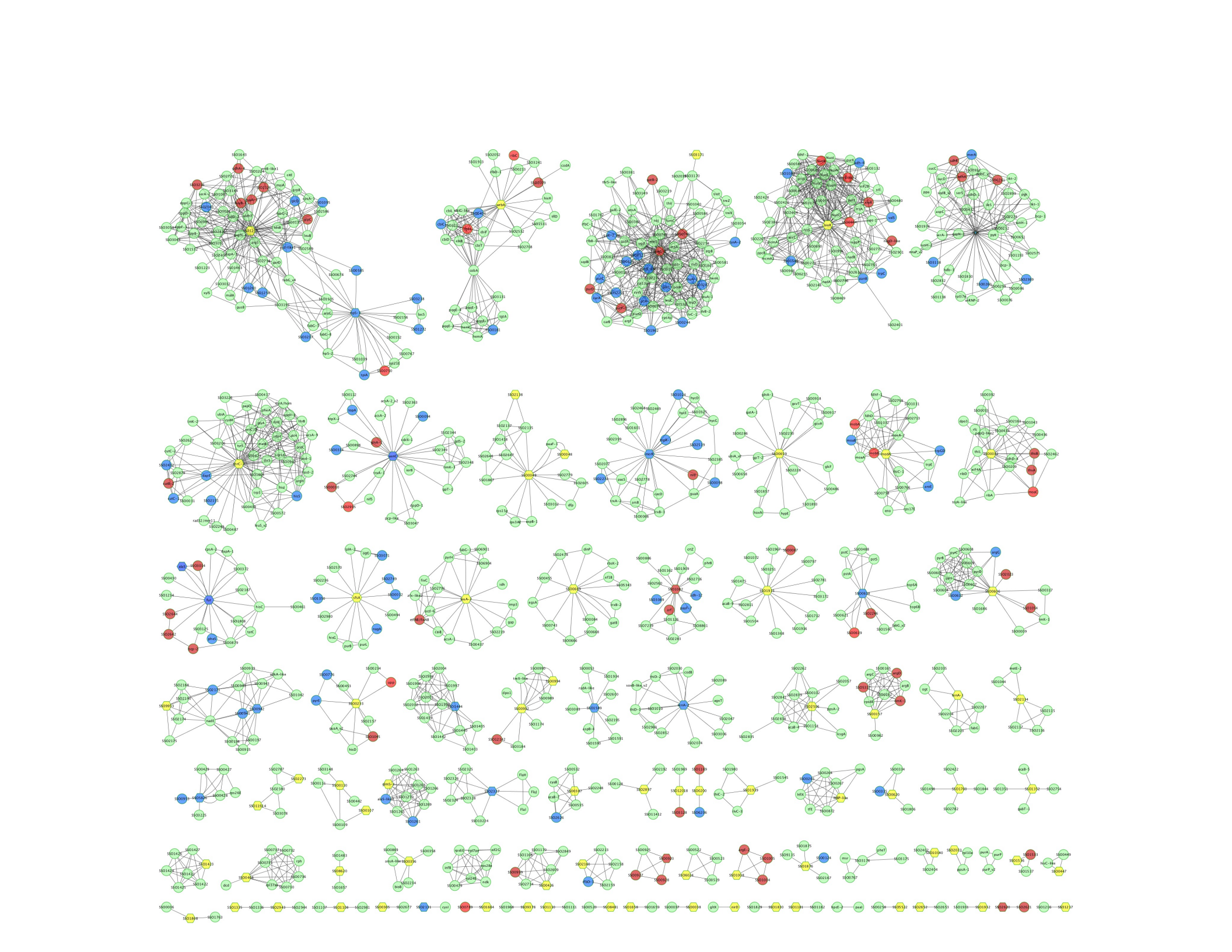
Sulfolobus solfataricus

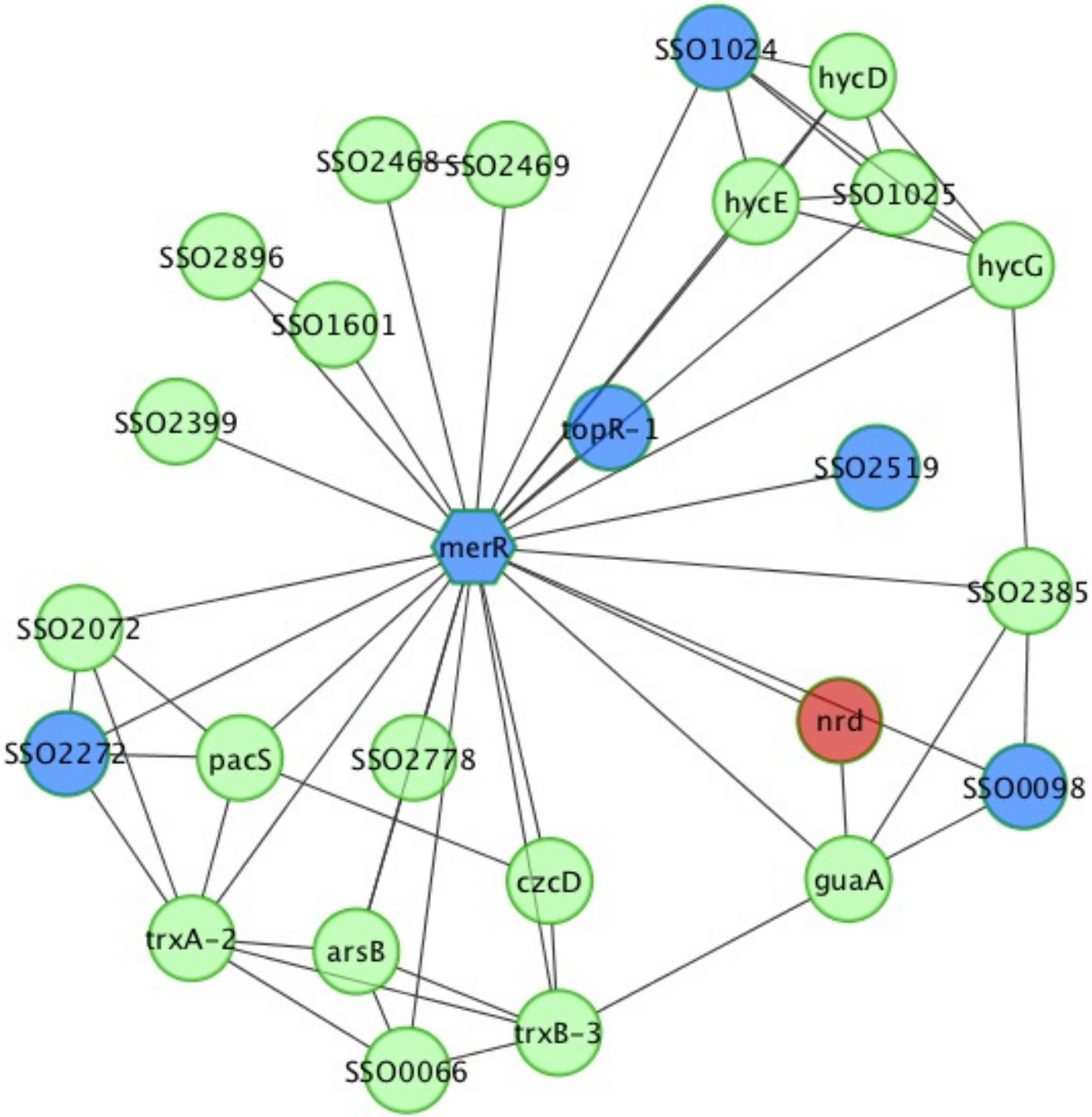
Crenarchaeota

Hyperthermophilic acidophilic sulfur-metabolizing archeon

2977 ORFs

111 FTs (3.2%)




SSO1024
hycD
SSO2468
SSO2469
hycE
SSO1025
hycG
SSO2896
SSO1601
SSO2399
topR-1
SSO2519
merR
SSO2385
SSO2072
nrd
SSO2272
pacS
SSO2778
SSO0098
czcD
guaA
trxA-2
arsB
trxB-3
SSO0066

Conclusions

Archaeal regulators

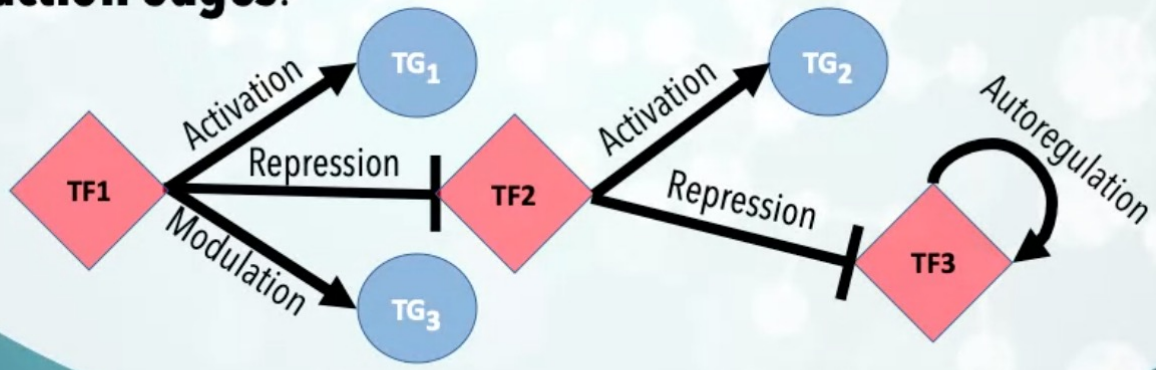
- bacterial-like
- Small sizes (lenght size) in contrast to bacteria
- Few tfs per genome:
- Multiple complexes
- Inference of regulatory networks in *S. solfataricus*

Similarities in biology

Gene Regulatory Network (GRN)

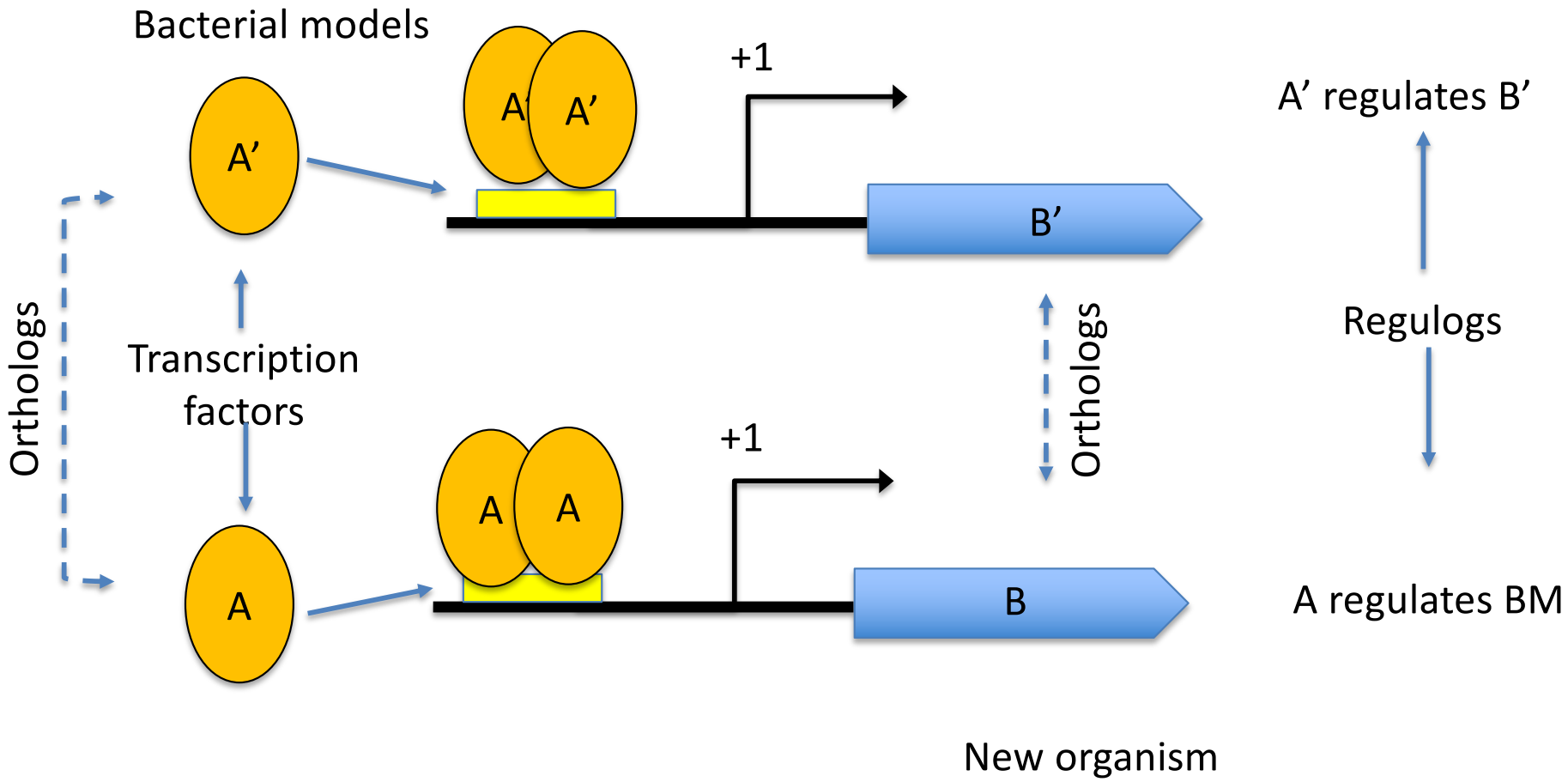
- **Regulatory interaction** is defined as a **TF-TG relation**
  - **TF** is the regulator gene [coding and noncoding (miRNAs, lncRNAs, etc)]
  - **TG** is the target gene

- A **GRN** is a **directed graph** in which **TFs** are connected to **TGs** by **interaction edges**.



(Karlebach and Shamir, 2008)

Inference of Gene Regulatory networks

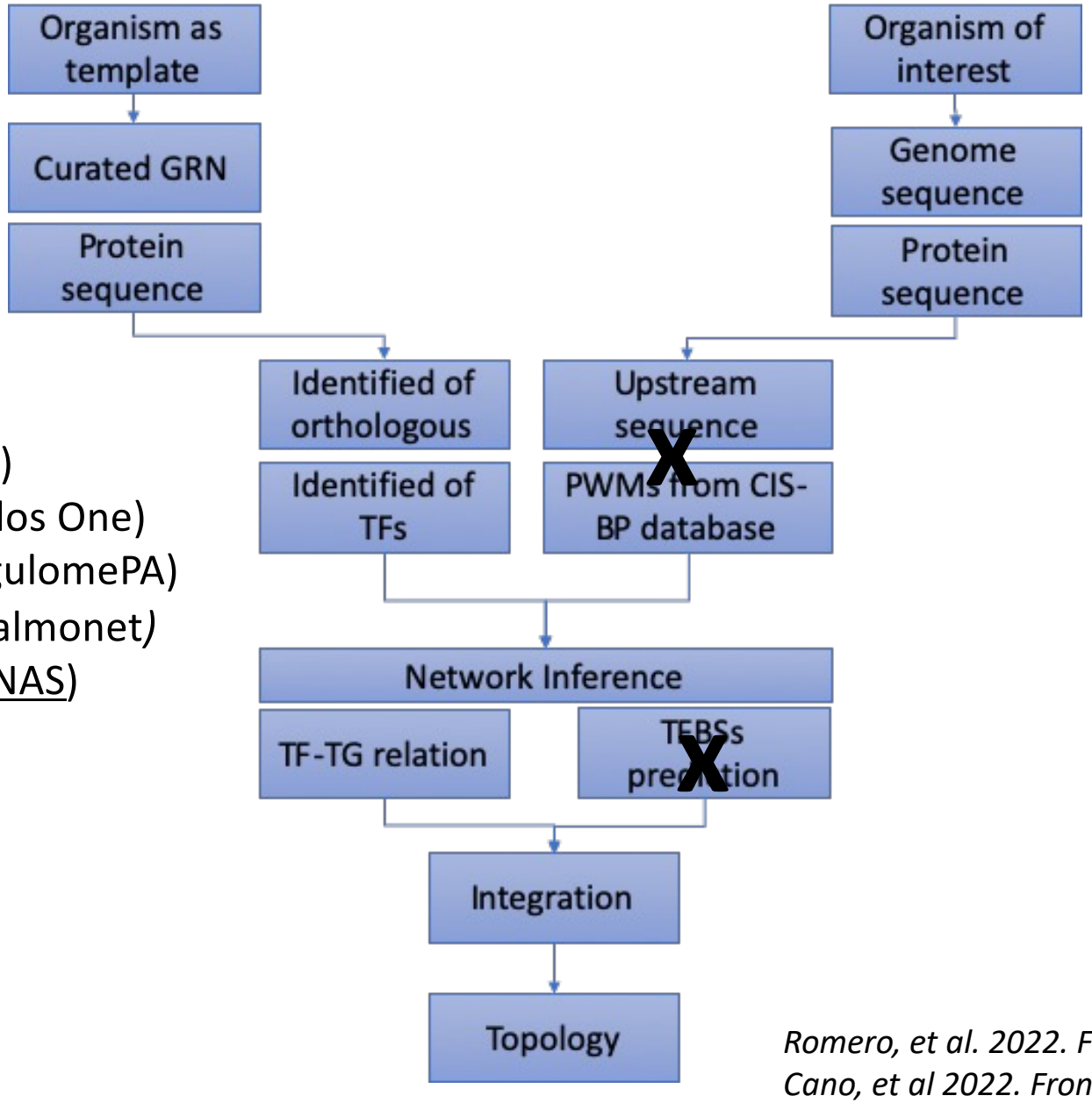


Romero, et al. 2022. *Front. Microbiol*

Inference of Gene Regulatory networks

Bacterial models

B. subtilis (DBTBS)
E. coli (RegulonDB)
M. tuberculosis (Plos One)
P. aeruginosa (RegulomePA)
*S. typhimurium (*Salmonet*)*
S. aureus N315 (PNAS)



Romero, et al. 2022. Front. Microbiol
Cano, et al 2022. Front. Microbiol
Galan-Vasquez, et al. 2023. PlosOne

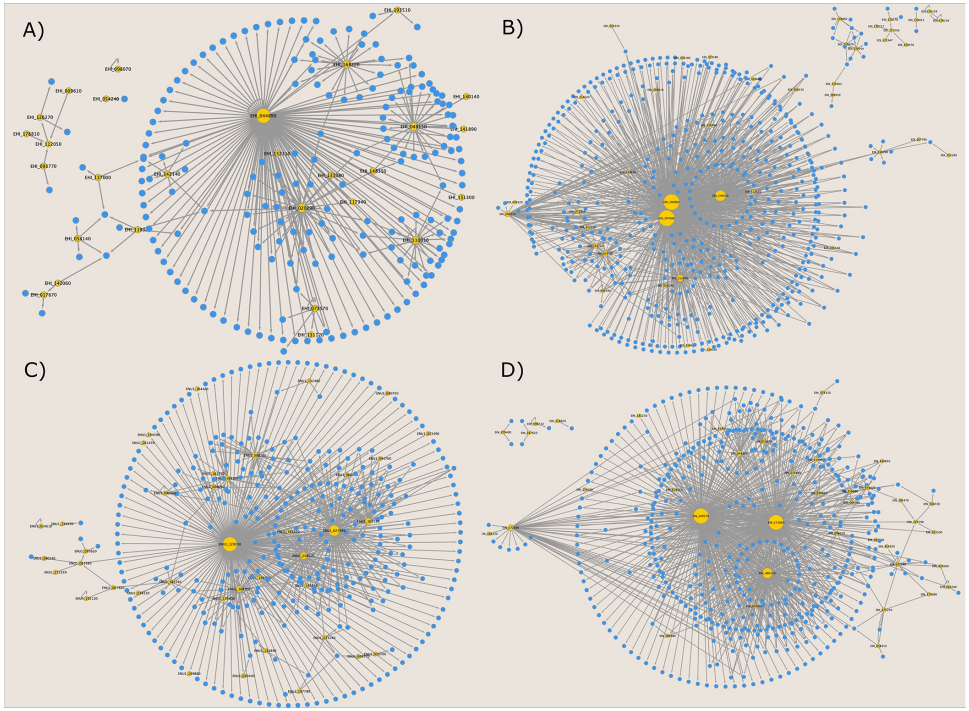
Inference of Gene Regulatory networks

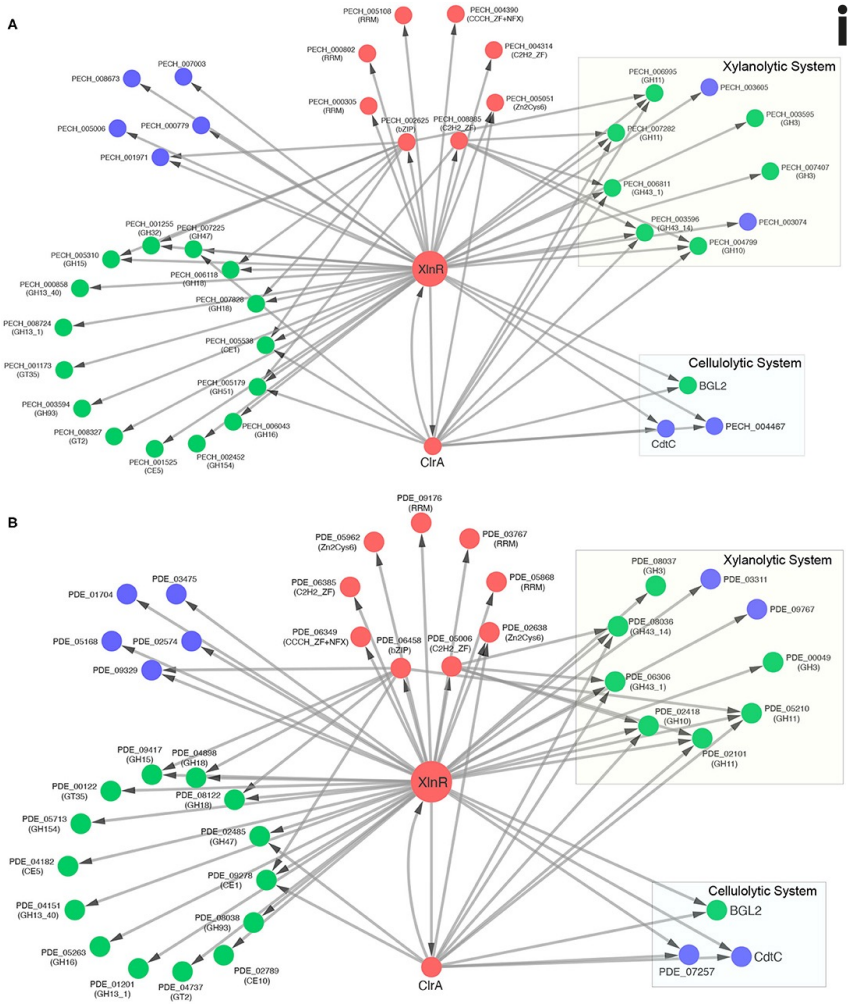
Bacterial model	TFs Original	TFs Extended	Targets Original	Targets Extended	Edge Original	Edge Extended
B_subtilis	191	229	1748	2145	2741	4166
E_coli_K12	196	252	1618	2145	3808	5138
P_aeruginosa_PA01	124	236	604	1609	1159	3315
S_enterica_LT2	131	225	1640	2282	2973	5187
S_aureus_N315	51	101	584	973	746	1538
M_tuberculosis_H37Rv	76	107	1405	1629	2637	3240

We found an increase in the number of targets, TFs, nodes and interactions for all the bacterial extended networks.
For instance, for *M. tuberculosis H37Rv* there is an increase of 224 TGs and 31 TFs; whereas in *B. subtilis*, 38 additional TFs and 397 new TGs.

Romero, et al. 2022. Front. Microbiol

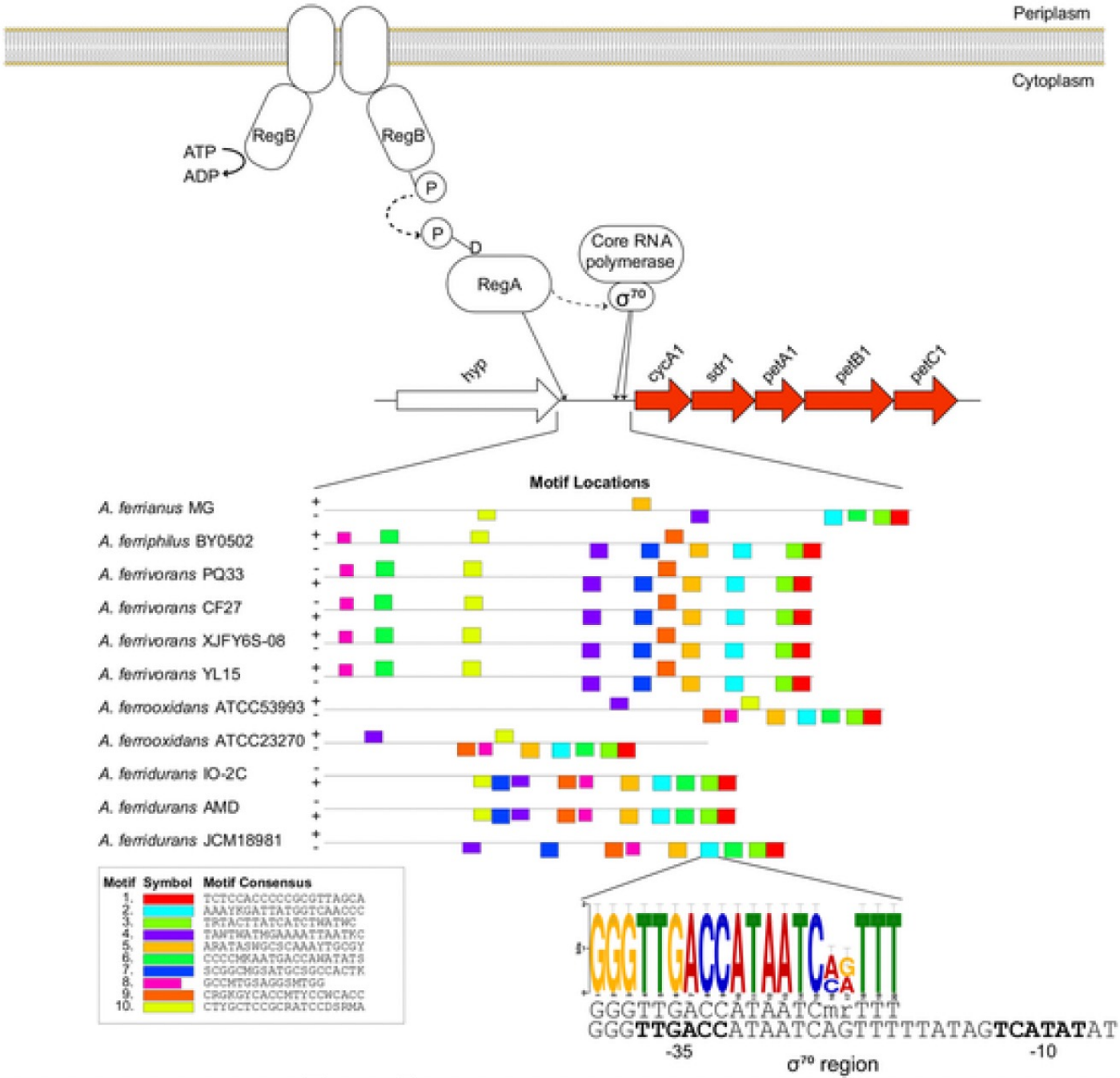
Entamoeba histolytica





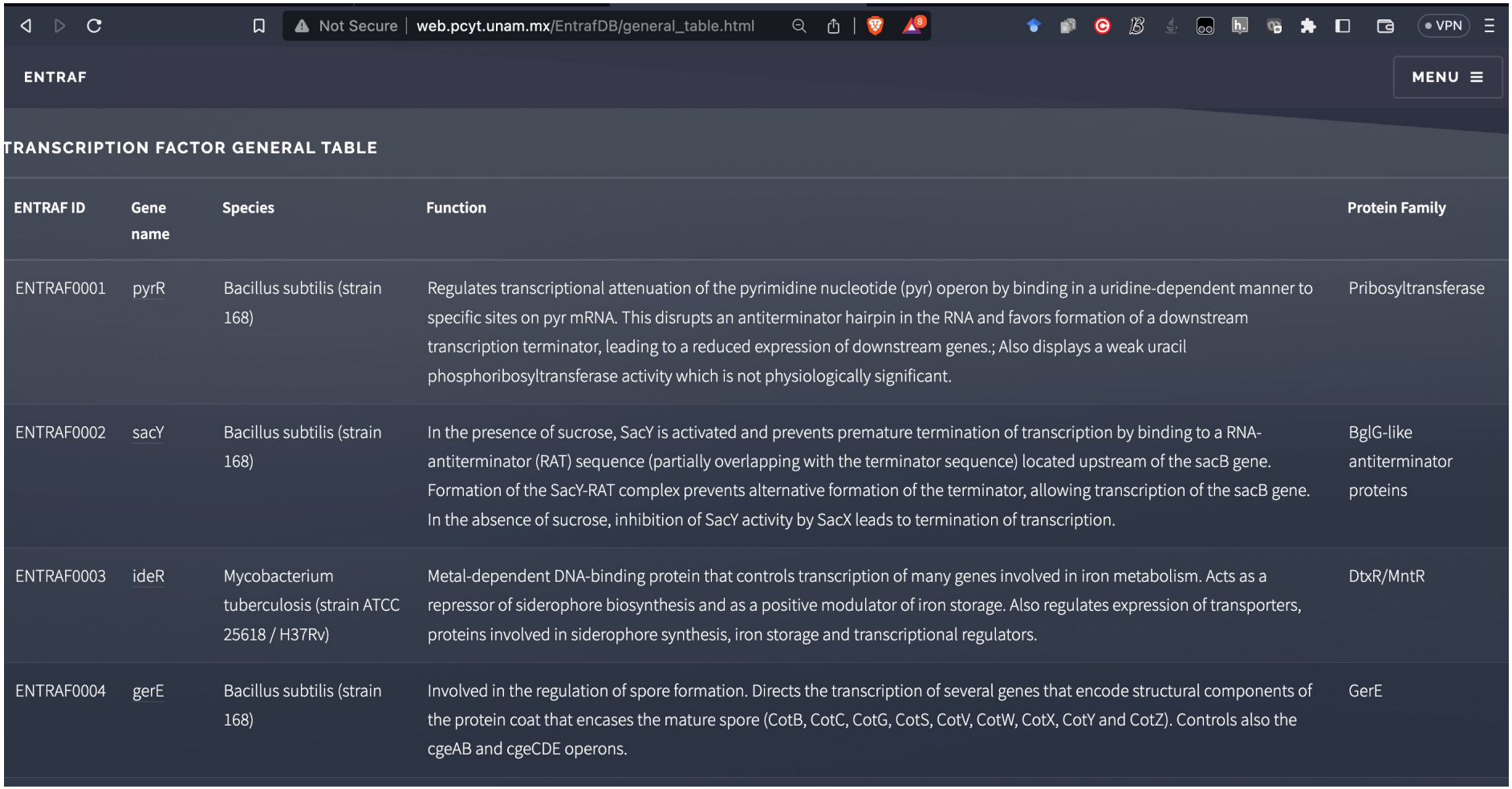
GRNs of A) *E. histolytica*, B) *E. dispar*, C) *E. nuttalli*, and D) *E. invadens*. The yellow nodes are TFs and the blue nodes are TGs; size nodes are proportional to output degree.

Galan-Vasquez E, *et al.* **2022**. PLoS ONE 17(8): e0271640.
Cano R, *et al.* **2022**. Frontiers in Microbiology. 13:861528
Lenz AR, *et al.* **2020**. Frontiers in Microbiology. 11:588263



Sepulveda, et al 2024. *Research in Microbiology*

ENTRAF

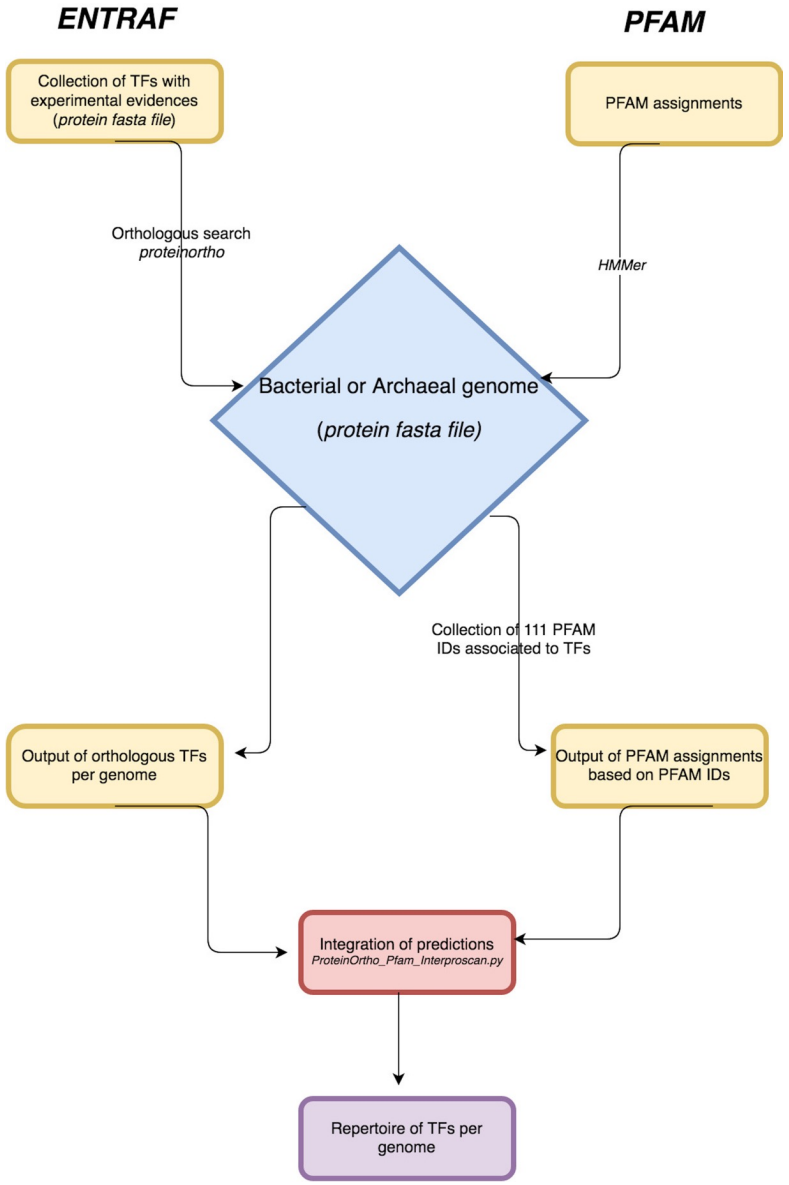


TRANSCRIPTION FACTOR GENERAL TABLE

ENTRAF ID	Gene name	Species	Function	Protein Family
ENTRAF0001	pyrR	Bacillus subtilis (strain 168)	Regulates transcriptional attenuation of the pyrimidine nucleotide (pyr) operon by binding in a uridine-dependent manner to specific sites on pyr mRNA. This disrupts an antiterminator hairpin in the RNA and favors formation of a downstream transcription terminator, leading to a reduced expression of downstream genes.; Also displays a weak uracil phosphoribosyltransferase activity which is not physiologically significant.	Pribosyltransferase
ENTRAF0002	sacY	Bacillus subtilis (strain 168)	In the presence of sucrose, SacY is activated and prevents premature termination of transcription by binding to a RNA-antiterminator (RAT) sequence (partially overlapping with the terminator sequence) located upstream of the sacB gene. Formation of the SacY-RAT complex prevents alternative formation of the terminator, allowing transcription of the sacB gene. In the absence of sucrose, inhibition of SacY activity by SacX leads to termination of transcription.	BglG-like antiterminator proteins
ENTRAF0003	ideR	Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv)	Metal-dependent DNA-binding protein that controls transcription of many genes involved in iron metabolism. Acts as a repressor of siderophore biosynthesis and as a positive modulator of iron storage. Also regulates expression of transporters, proteins involved in siderophore synthesis, iron storage and transcriptional regulators.	DtxR/MntR
ENTRAF0004	gerE	Bacillus subtilis (strain 168)	Involved in the regulation of spore formation. Directs the transcription of several genes that encode structural components of the protein coat that encases the mature spore (CotB, CotC, CotG, CotS, CotV, CotW, CotX, CotY and CotZ). Controls also the cgeAB and cgeCDE operons.	GerE

Flores-Bautista E, *et al*. 2020. PLoS One
Martinez-Liu L, *et al*. 2021. PLoS One.

Prediction of TFs



Entraf: A collection of 668 TFs with experimental evidences.

111 PFAM IDs associated to TFs

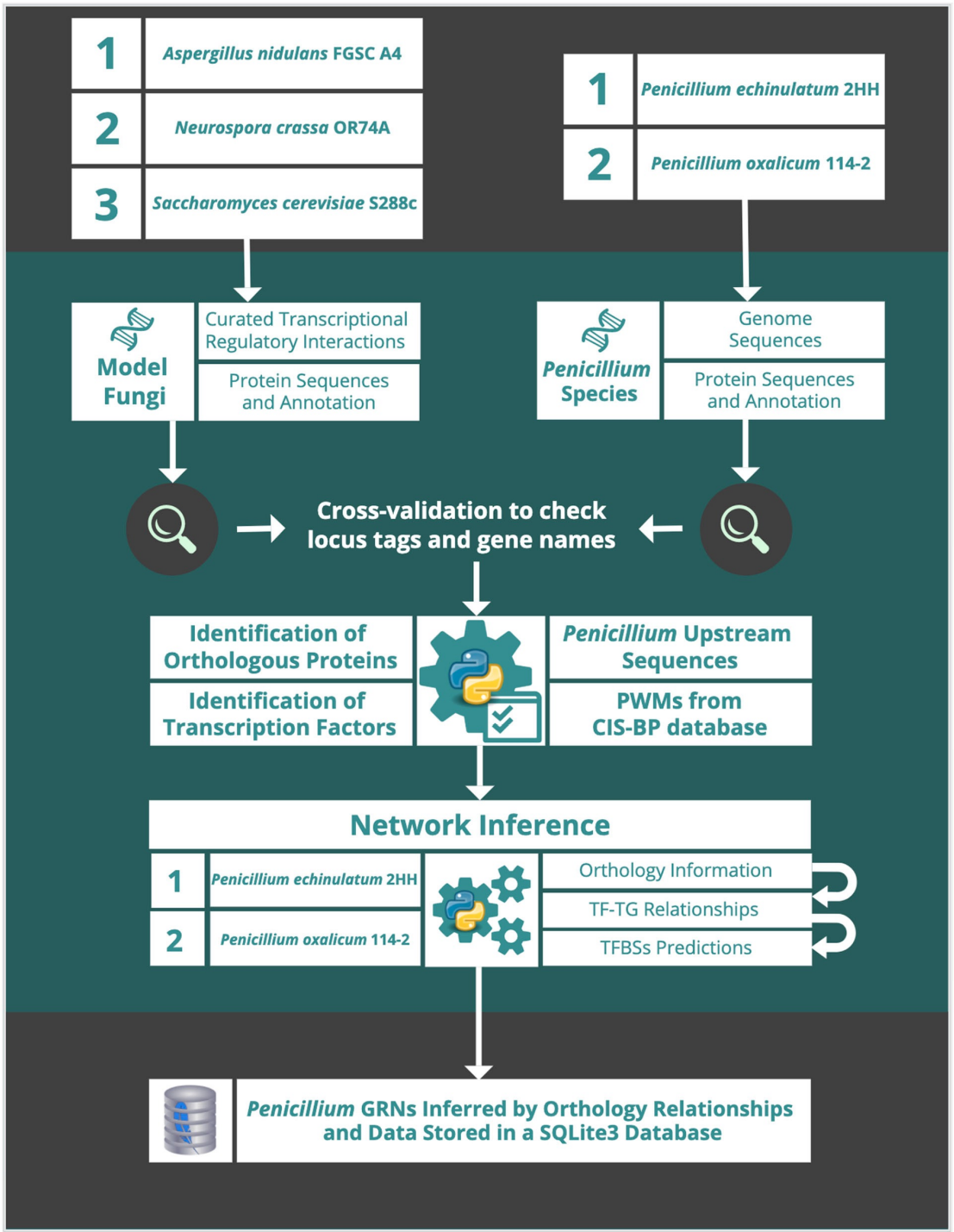
- *DBD*
- *RegulonDB*
- *DBTBS*
- *CollecTF*
- *Entraf*

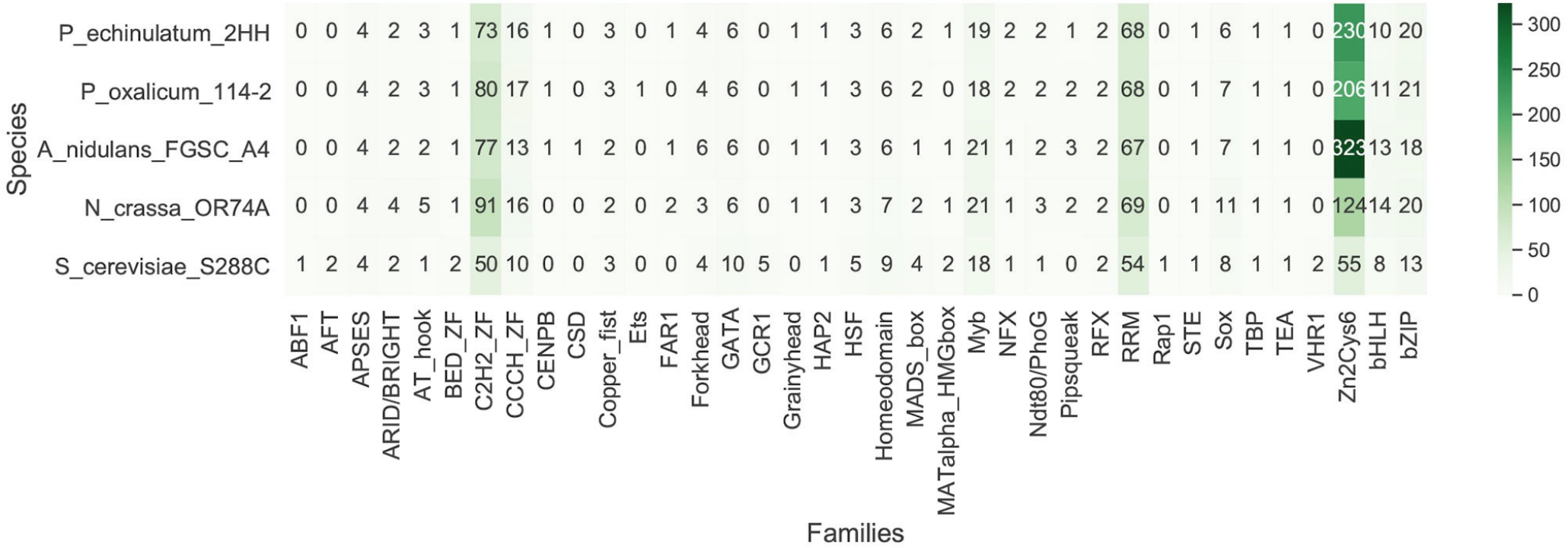
http://pcyt.unam.mx/EntrafDB/

Ledesma L, et al. 2024. Scientific Reports.
Ledesma L, et al. 2022. In Prokaryotic Gene Regulation: Methods in Molecular Biology.

Gene Regulatory Networks of *Penicillium echinulatum* 2HH and *Penicillium oxalicum* 114-2 Inferred by a Computational Biology Approach

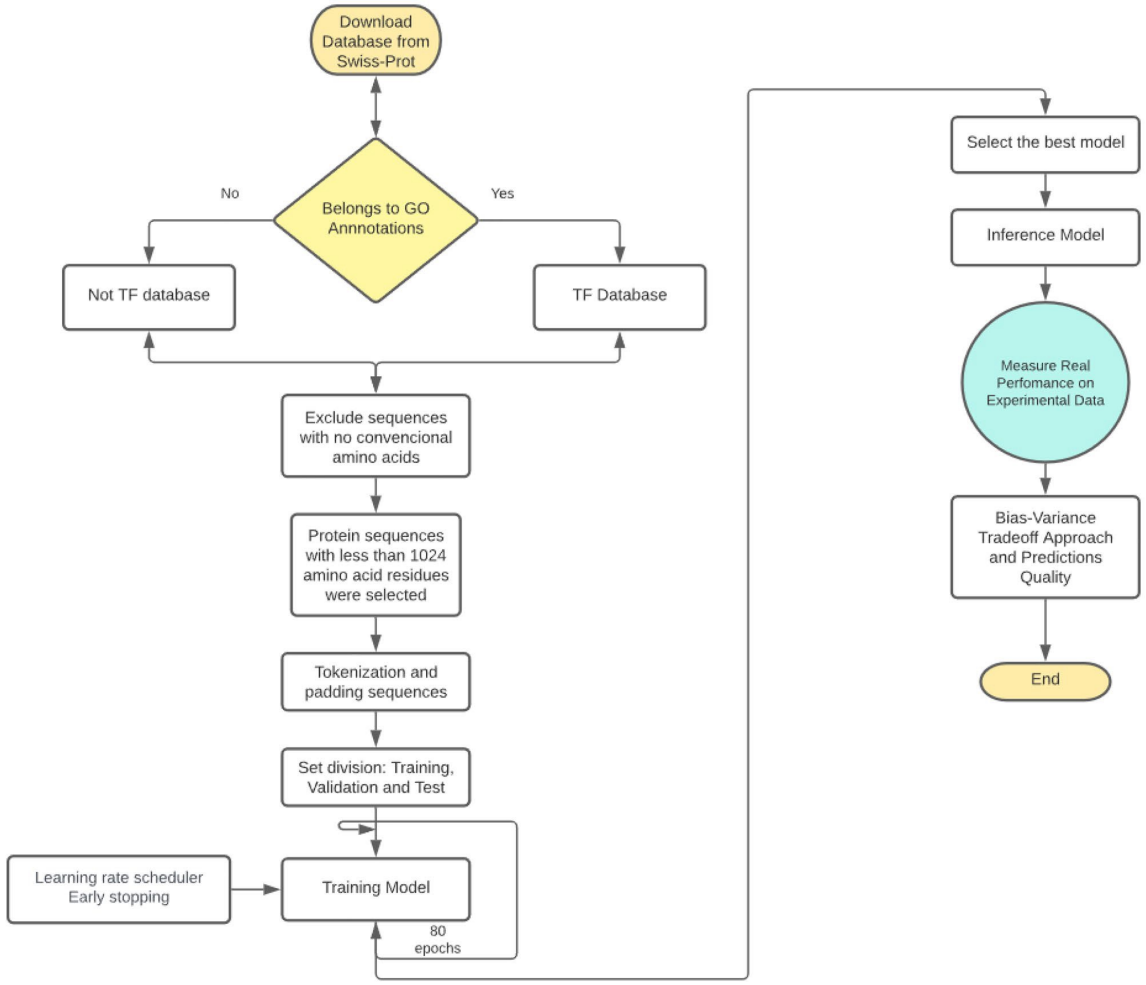
Alexandre Rafael Lenz[1,2,3], Edgardo Galán-Vásquez[4], Eduardo Balbinot[2], Fernanda Pessi de Abreu[2], Nikael Souza de Oliveira[2,5], Letícia Osório da Rosa[5], Scheila de Avila e Silva[2], Marli Camassola[5], Aldo José Pinheiro Dillon[5] and Ernesto Perez-Rueda[1,6*]*



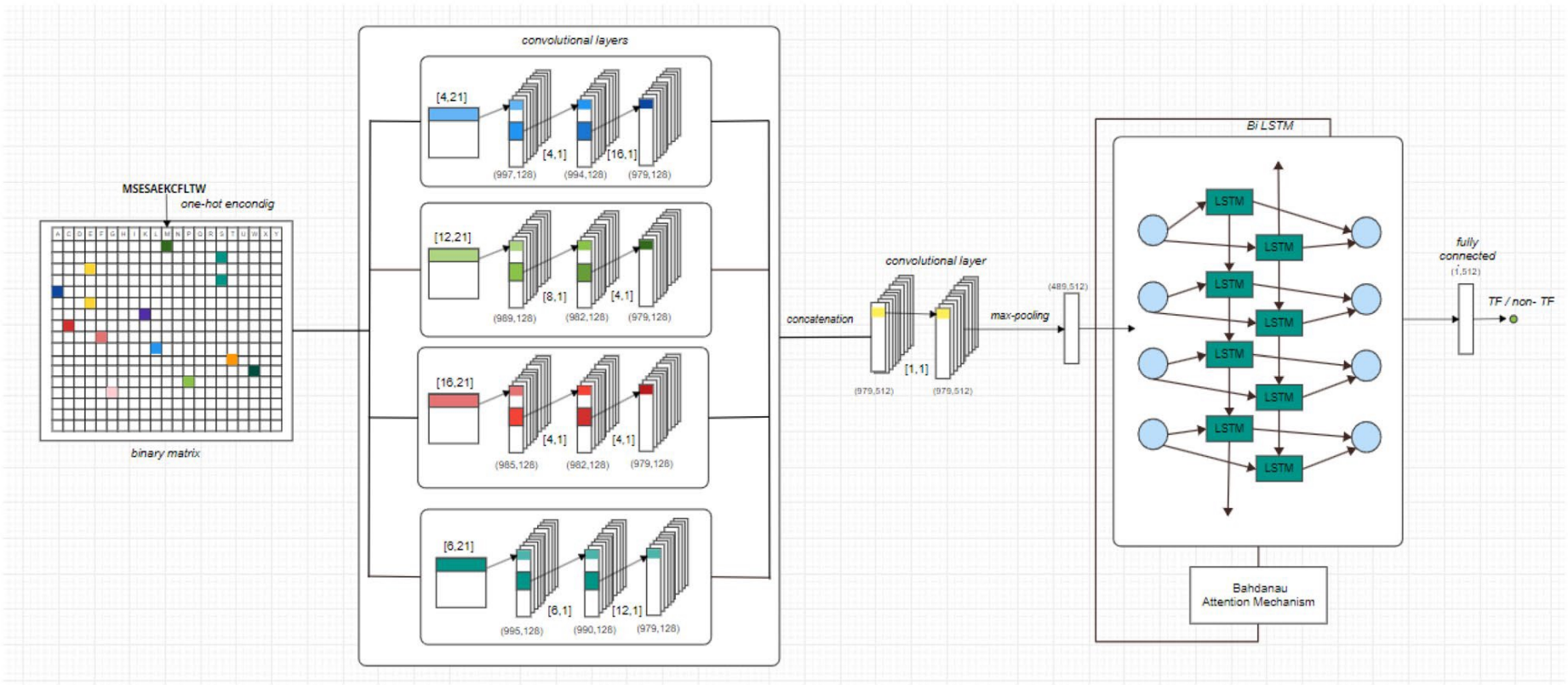


Abundance and distribution of transcription factor families in fungal genomes. TFs of *P. echinulatum* 2HH containing multiple domains: Zn2Cys6+C2H2_ZF (5); Homeodomain+C2H2_ZF (1); BED_ZF+C2H2_ZF (1); STE+C2H2_ZF (1); CCCH_ZF+NFX (1); RRM+CCCH_ZF (3); RRM+Zn2Cys6 (1). TFs of *P. oxalicum* 114-2 containing multiple domains: Zn2Cys6+C2H2_ZF (7); Homeodomain+C2H2_ZF (1); BED_ZF+C2H2_ZF (1); STE+C2H2_ZF (1); CCCH_ZF+NFX (1); RRM+CCCH_ZF (3).

DeepReg

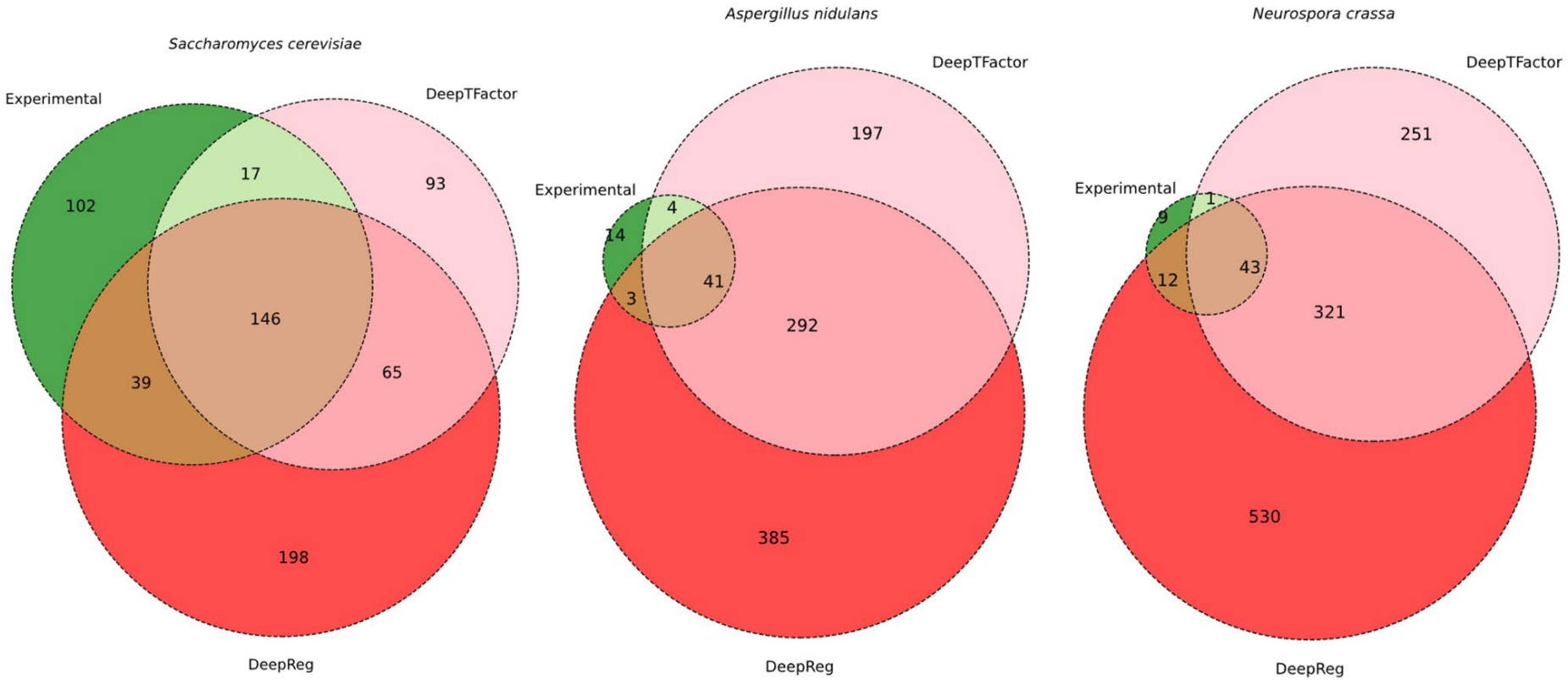


Flow diagram of the DL used to predict transcription factors. A total of 22,100 TFs and 527,146 non- TFs were retrieved from SwissProt. These sequences were cleaned to be tokenized and padded. In a posterior step, the training, validation, and test datasets were built at a ratio of 90:9:1, in that order. The model was trained for a maximum of 80 epochs using a learning rate scheduler and early stopping to avoid overfitting. After many trial and error rounds; changing hyperparameters, such as batch size, dropout rates, and initial learning rates. Finally, an inference model was produced to evaluate performance and quality against experimental data using a bias-variance tradeoff



The network architecture of DeepReg. Three modules were considered: (1) 4 CNN layers as the feature extractor, (2) LSTM, and (3) the attention mechanism module. Each sequence is tokenized and passes through a one-hot encoding process.

Efficiency of predictions associated with DeepReg and DeepTFactor. The dataset of TFs was predicted in three fungal models: S. cerevisiae, N. crassa, and A. nidulans, and compared against their repertoire of experimentally characterized TFs



- *Data availability*
- *The datasets analyzed during the current study are available in the Uniprot database, (https:// www. unipr ot. org/) and in the github https:// github. com/ Leona rdoLed/ DeepL earni ng-_ TF*

Acknowledgments

Dr. Andre Farias → Modelaje estructural y dinámica molecular

C. Dr. Daniel Cortes → Complejo estructural de regulación.

Dra. Dulce Alvarez → Integración redes regulatorias y de coexpresión de bacterias.

Dr. Jose Vllalpando → Metabolismo (bloques y repertorio enzimático)

Dra. Silvia Tenorio → Metabolismo y microbiomas

Dra. Marisa Fabiana --> Gene regulation

Dr. Leonardo Ledesma → Predicción de TFs con deep learning

Dr. Daniela Mares → Redes de coexpresión de cancer de mama

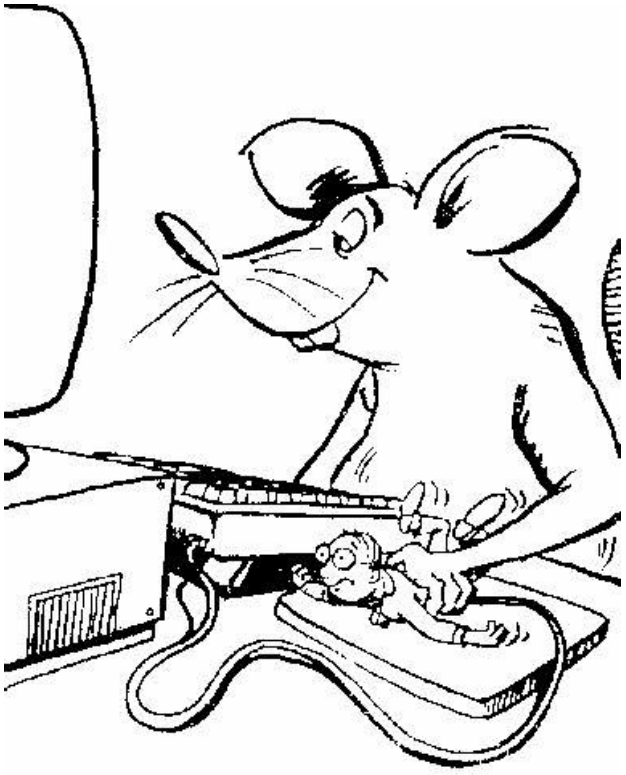
Dr. Edgardo Galan → WGCNA

Dr. Mario Alberto Martinez → Microbiomas costeros

Dr. Gustavo Martinez → ML & Promoters

CONAHCYT DGAPA CYTED

Thank you for your attention



http://openwetware.org/wiki/User:Ernesto_Perez-Rueda