RegulonDB: Enhancing access to genomic knowledge in *Escherichia coli* K-12

Julio Collado

Centro de Ciencias Genómicas
UNAM, Cuernavaca

30 years of TRANSFAC

Göttingen University, Germany
March 7th, 2018

Centro de Ciencias Genómicas

The challenge:

We are living a tsunami of data, information and knowledge

Can we make use of the knowledge that is being generated?

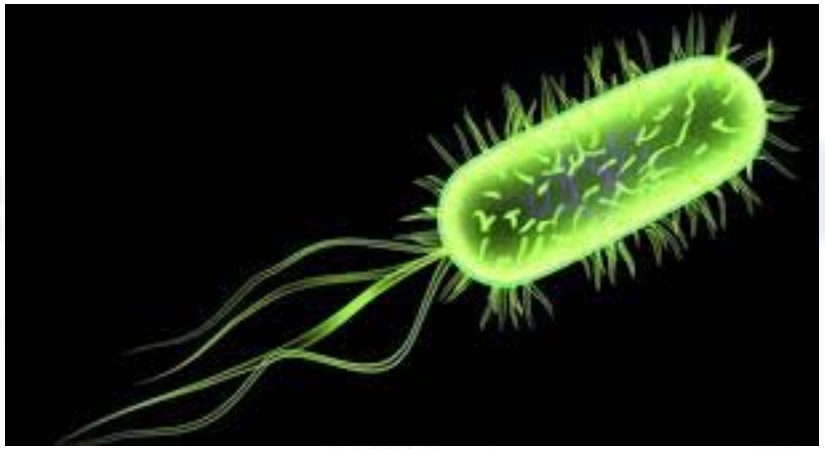
Isaac Asimov predicted the internet of today 20 years ago
https://www.youtube.com/watch?v=8ZmFEFO72gA

Gene regulation in *Escherichia coli* K-12

Table 4. Distribution of *E. coli* proteins among 22 functional groups (simplified schema).

Functional class	Number	Percent of total
Regulatory function	45	1.05
Putative regulatory proteins	133	3.10
Cell structure	182	4.24
Putative membrane proteins	13	0.30
Putative structural proteins	42	0.98
Phage, transposons, plasmids	87	2.03
Transport and binding proteins	281	6.55
Putative transport proteins	146	3.40
Energy metabolism	243	5.67
DNA replication, recombination, modification, and repair	115	2.68
Transcription, RNA synthesis, metabolism, and modification	55	1.28
Translation, posttranslational protein modification	182	4.24
Cell processes (including adaptation, protection)	188	4.38
Biosynthesis of cofactors, prosthetic groups, and carriers	103	2.40
Putative chaperones	9	0.21
Nucleotide biosynthesis and metabolism	58	1.35
Amino acid biosynthesis and metabolism	131	3.06
Fatty acid and phospholipid metabolism	48	1.12
Carbon compound catabolism	130	3.0
Central intermediary metabolism	188	4.3
Putative enzymes	251	5.8
Other known genes (gene product or phenotype known)	26	0.6
Hypothetical, unclassified, unknown	1632	38.0
Total	4288	100.0

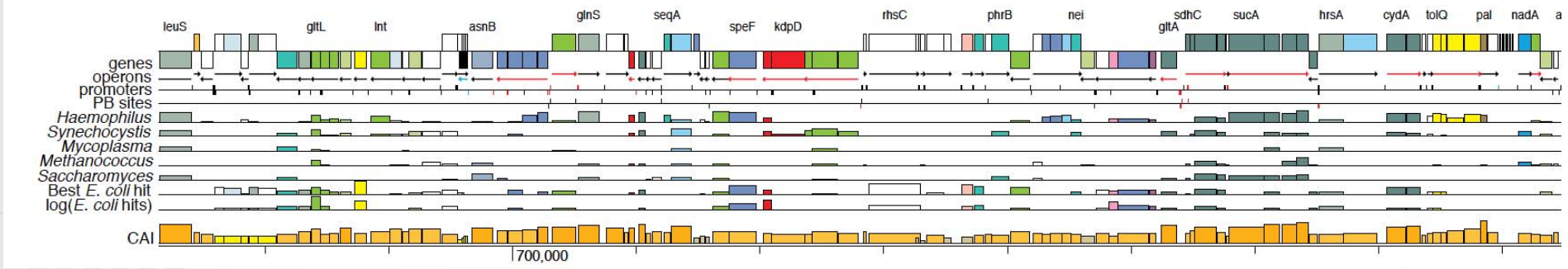
*Total of these rounded values is 99.97%.



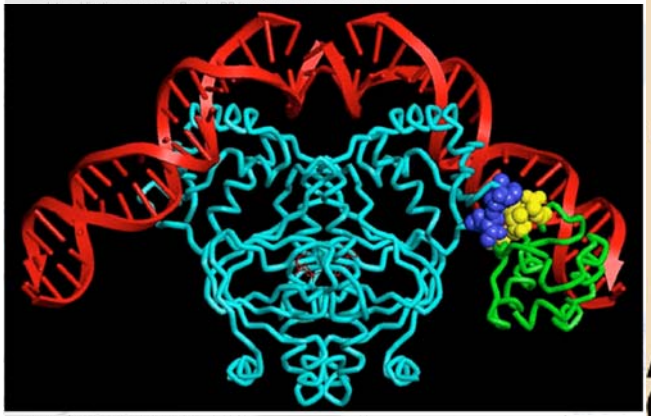
ARTICLE

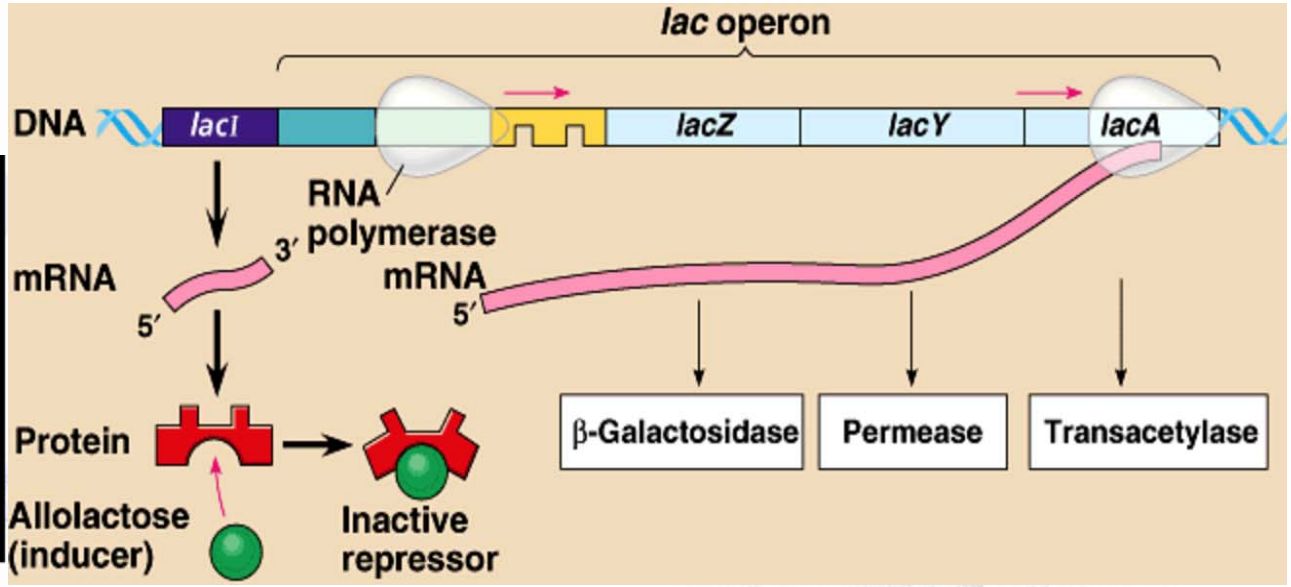
The Complete Genome Sequence of *Escherichia coli* K-12

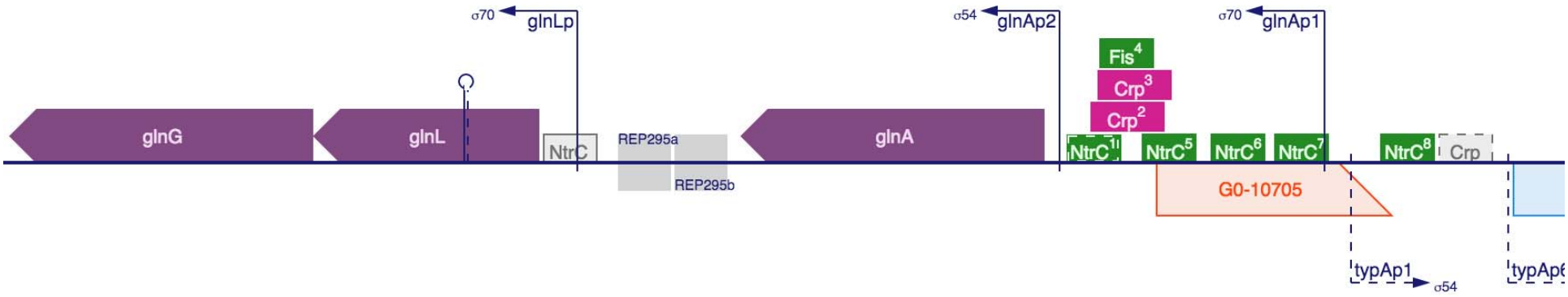
Frederick R. Blattner,* Guy Plunkett III,* Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau, Ying Shao



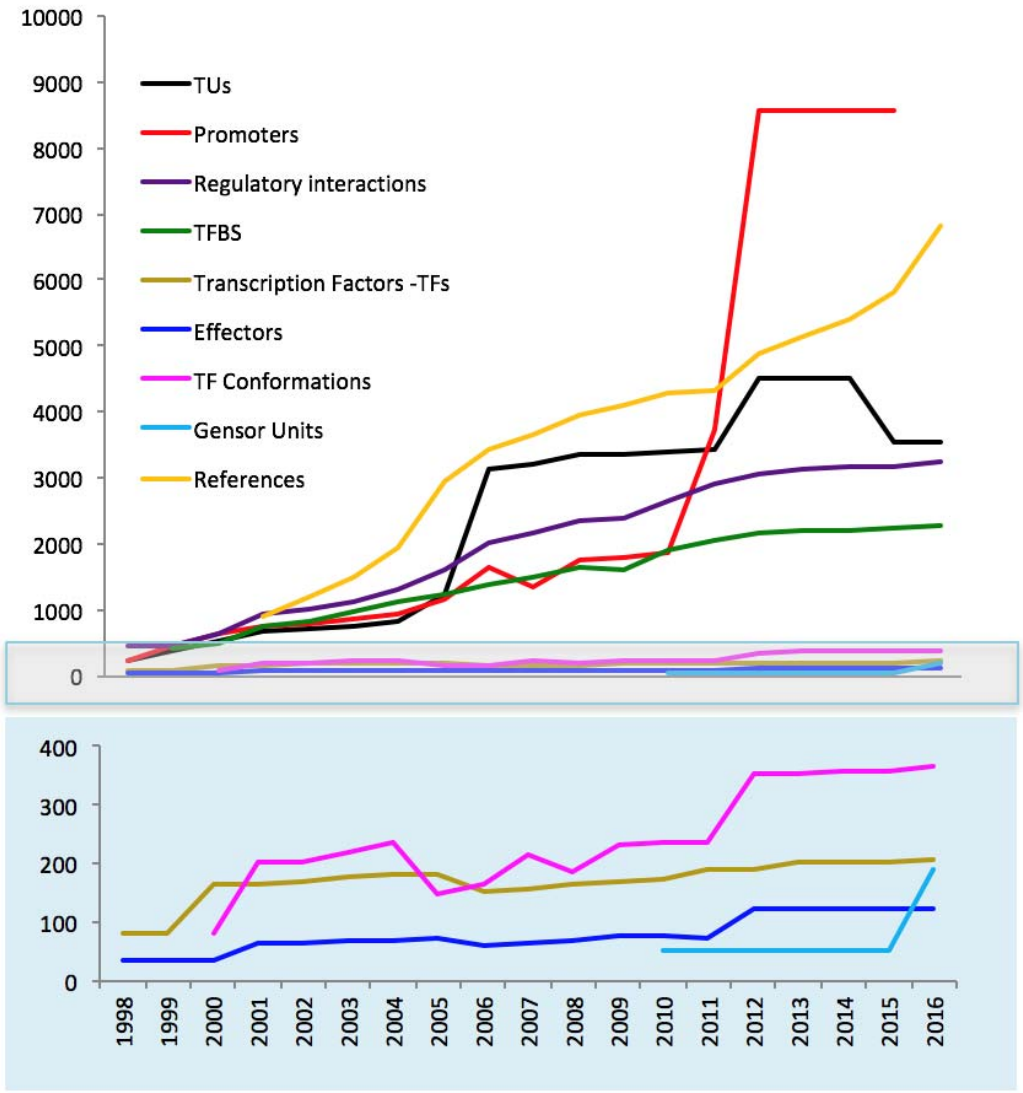
Promoters, operators, transcription factors, operons and transcription units



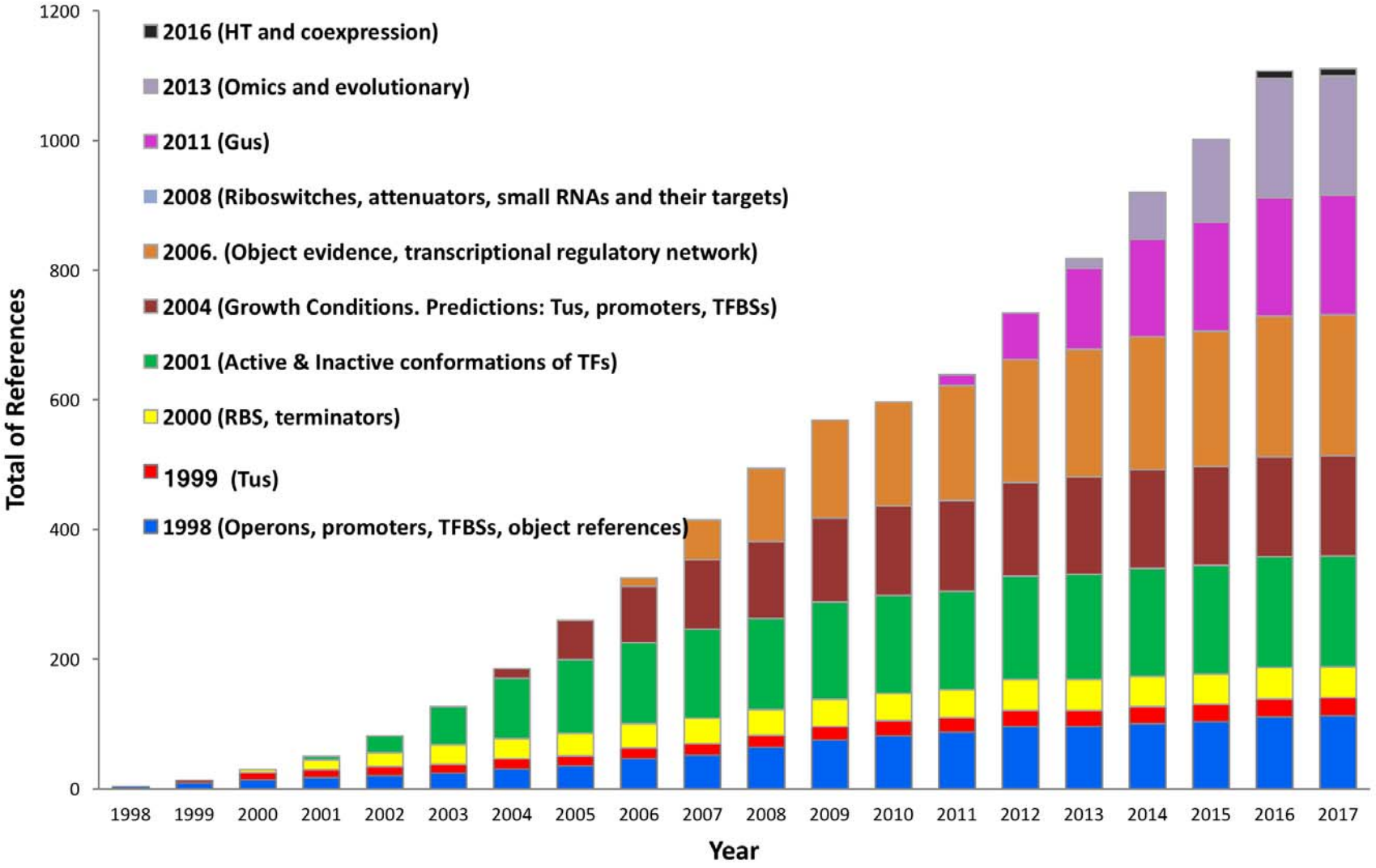




Curation through the years in RegulonDB



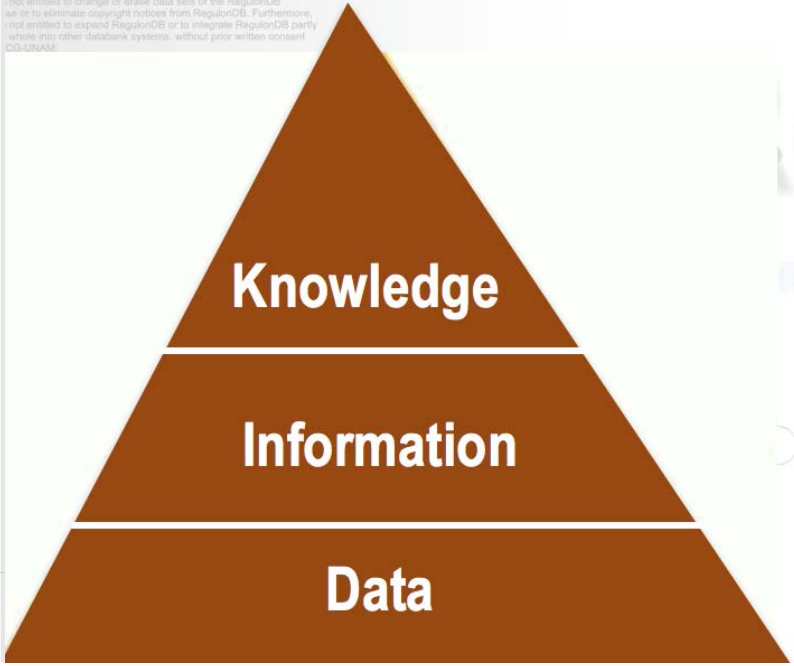
Papers and new objects through the years



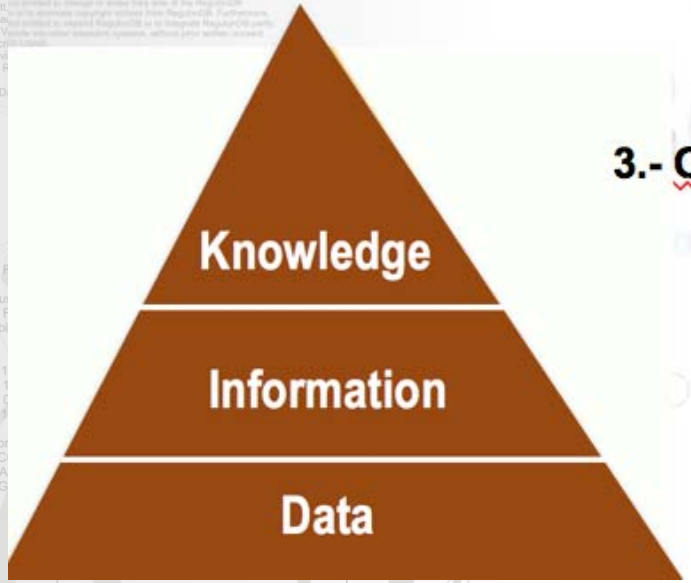
The main goal of biocuration:

To make data, information and knowledge easily accessible to humans (and computers)

I will illustrate 2 examples



2.- Operons, regulons and GENSOR Units

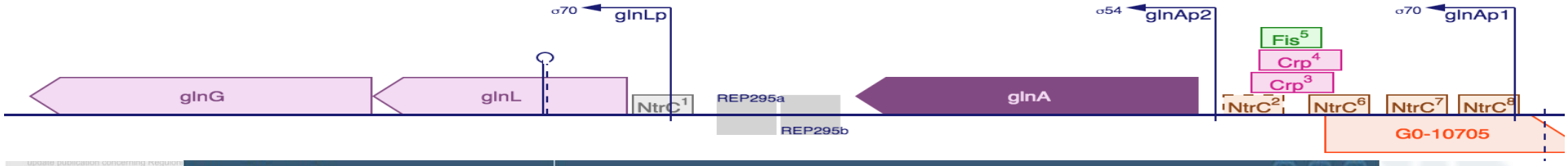


3.- Ontology and NLP strategies

2.- Operons, regulons and GUs

1.- Curating HT experiments

Understanding is a construction



Transcription unit

Name: glnALG

Promoter

Name: glnAp2
+1: 4058107
Sigma Factor: Sigma54 **Sigmulon**
Distance from start of the gene: 73
Sequence: gcatgataacgccttttaggggcaatttaaaagttggcacagatttcgcttttatcttttttacggcgacacggccaaaata
−24 −12 +1
Evidence: [TIM]
Reference(s):
[5] Ninfa AJ., et al., 1987
[6] Reitzer LJ., et al., 1989
[7] Tian ZX., et al., 2001

TF binding sites (TFBSs)

[CRP,-] Phrase

Type	Transcription factor	Function	Promoter	Binding Sites				Growth Conditions	Evidence (Confirmed, Strong, Weak)	Reference(s)
				LeftPos	RightPos	Central Rel-Pos	Sequence			
proximal	CRP-cAMP	repressor	glnAp2	4058152	4058173	-55.5	ttttgcacgaTGGTGCGCATGATAACGCCTTTtaggggcaat	nd	[AIBSCS], [BPP], [GEA]	[7]
proximal	CRP-cAMP	repressor	glnAp2	4058154	4058175	-57.5	ccttttgcacGATGGTGCGCATGATAACGCCTtttaggggca	nd	[AIBSCS], [BPP], [GEA]	[7]

[Fis,+] Phrase

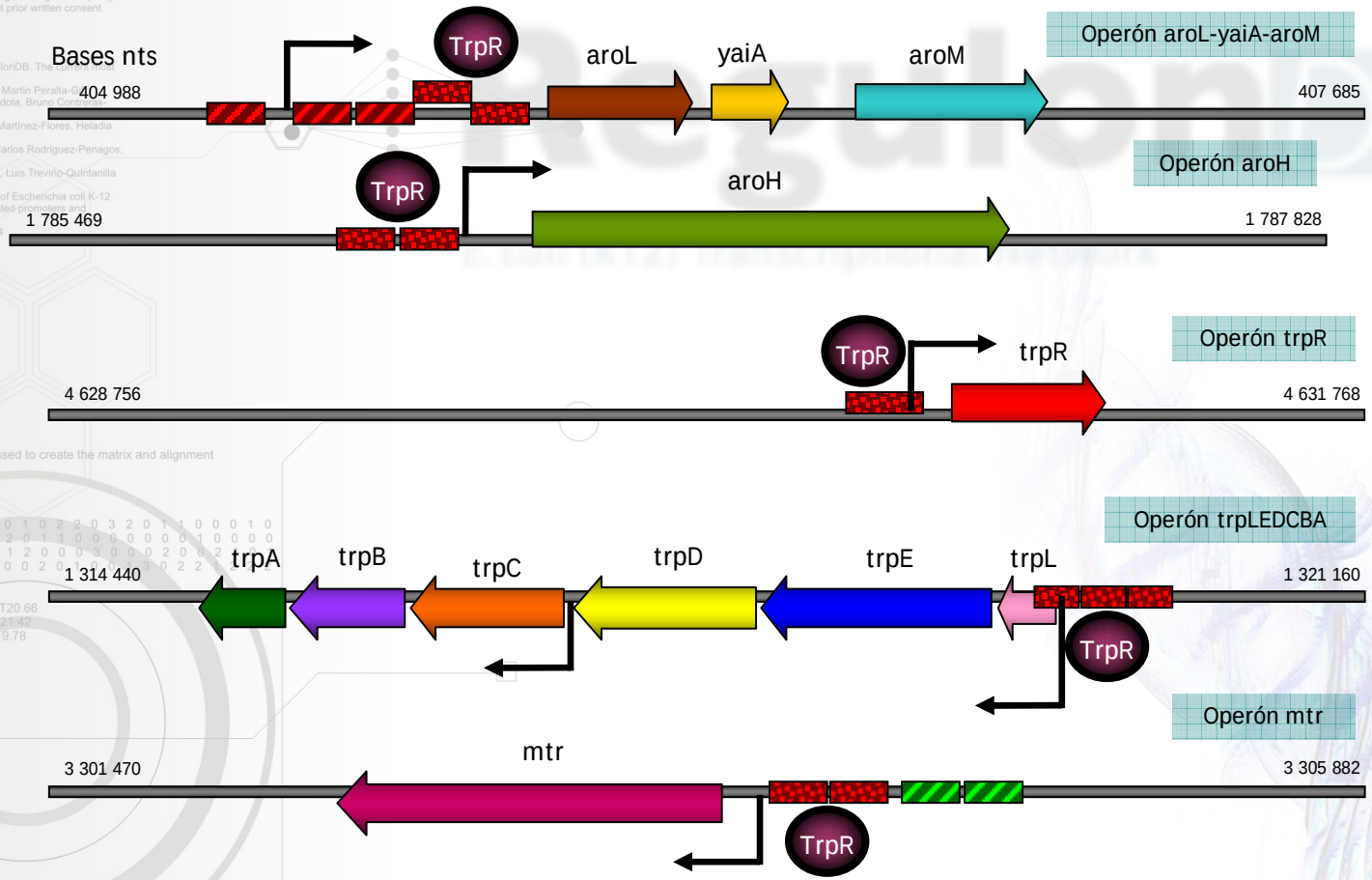
Type	Transcription factor	Function	Promoter	Binding Sites				Growth Conditions	Evidence (Confirmed, Strong, Weak)	Reference(s)
				LeftPos	RightPos	Central Rel-Pos	Sequence			
proximal	Fis	activator	glnAp2	4058155	4058169	-55.0	gcacgatggtGCGCATGATAACGCCttttaggggc	nd	[AIBSCS], [BPP], [GEA]	[12]

[NtrC,+] Phrase

Type	Transcription factor	Function	Promoter	Binding Sites				Growth Conditions	Evidence (Confirmed, Strong, Weak)	Reference(s)
				LeftPos	RightPos	Central Rel-Pos	Sequence			
proximal	NtrC-Phosphorylated	activator	glnAp2	4058144	4058160	-45.0	tgcgcatgatAACGCCTTTTAGGGGCAatttaaaagt	nd	[GEA], [HIBSCS]	[9], [11]
proximal	NtrC-Phosphorylated	activator	glnAp2	4058167	4058183	-68.0	ggtgcagcccTTTTGCACGATGGTGCGcatgataacg	nd	[GEA], [HIBSCS], [SM]	[9], [11]
proximal	NtrC-Phosphorylated	activator	glnAp2	4058188	4058204	-89.0	ttggtgcaacATTCACATCGTGGTGCAgccctttttgc	nd	[BPP], [GEA], [HIBSCS], [SM]	[5], [6], [8], [9], [11]
remote	NtrC-Phosphorylated	activator	glnAp2	4058207	4058223	-108.0	tttccattgaAGCACTATATTGGTGCAacattcacat	nd	[BPP], [GEA], [HIBSCS], [SM]	[5], [6], [9], [10]
remote	NtrC-Phosphorylated	activator	glnAp2	4058239	4058255	-140.0	caaaggtcatTGCACCAACATGGTGCTtaatgtttcc	nd	[BPP], [GEA], [HIBSCS], [SM]	[5], [6], [8], [9], [10]

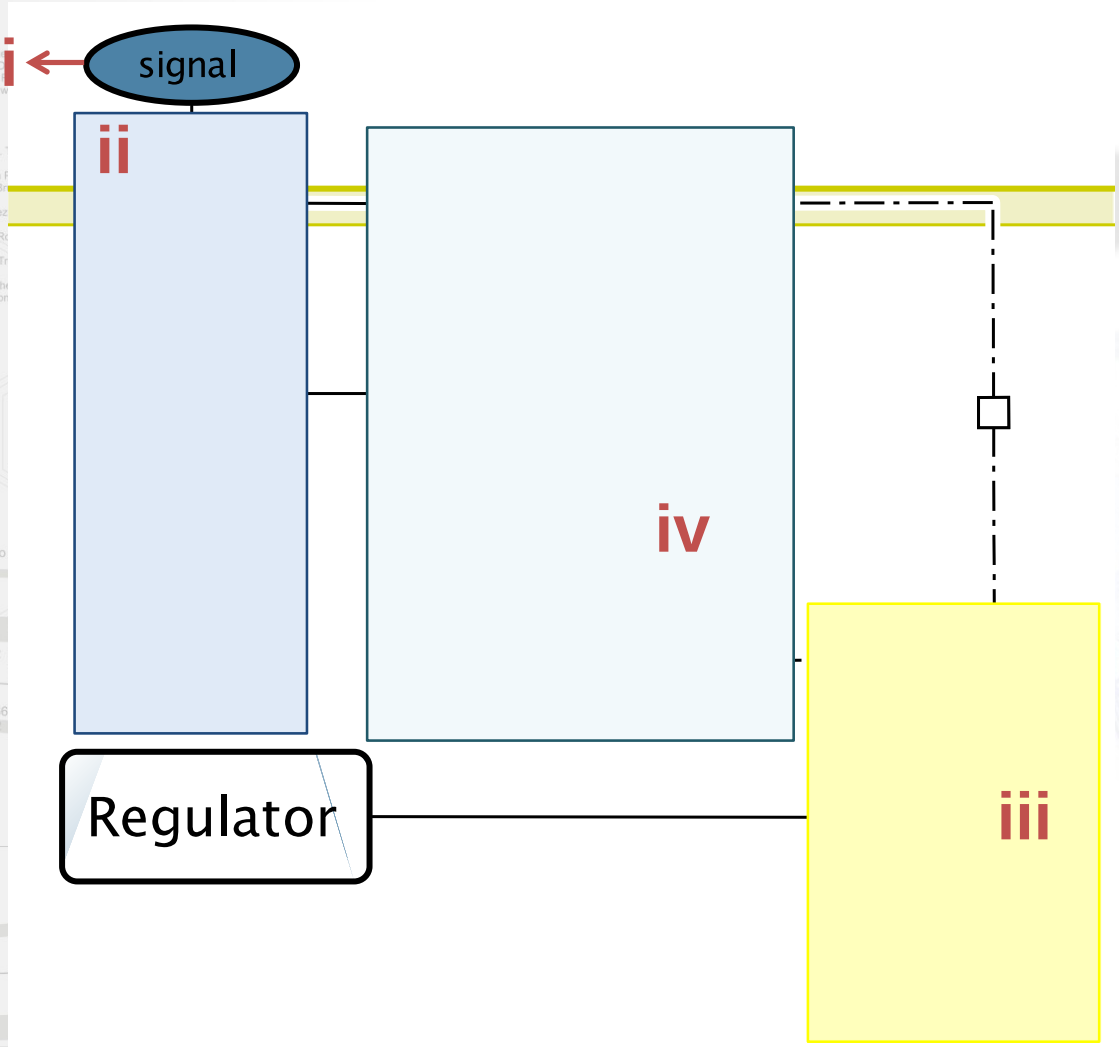
Groups of properties constitute objects, and some objects contain multiple objects

The tryptophan (TrpR) regulon



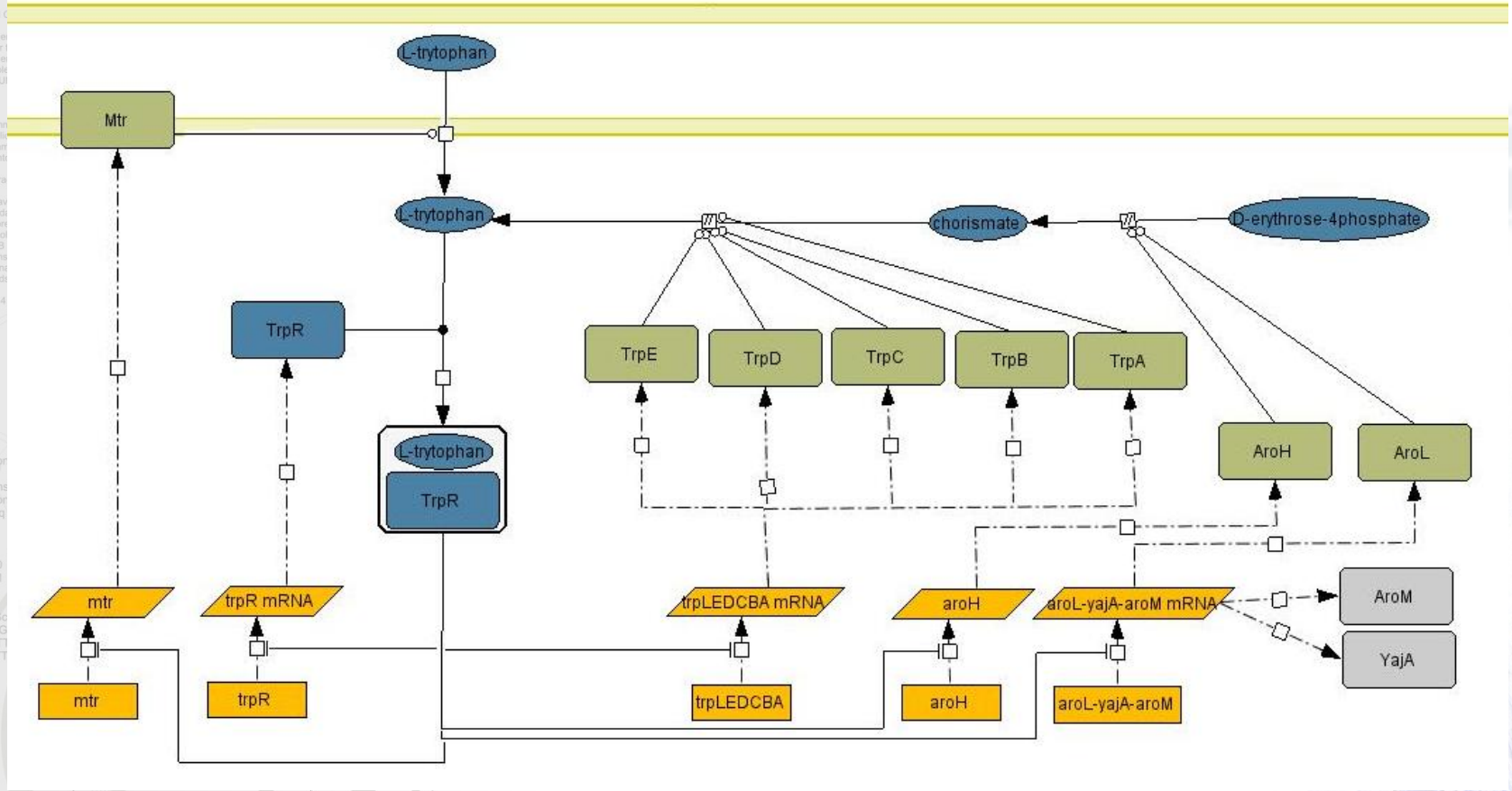
What is TrpR controlling ?

A new concept: GENSOR Units



DNA RNA Protein Molecule Complex Transition Reaction Catalysis
Secondary Reaction Transcription Translation Induction Inhibition Inner/Outer Membrane

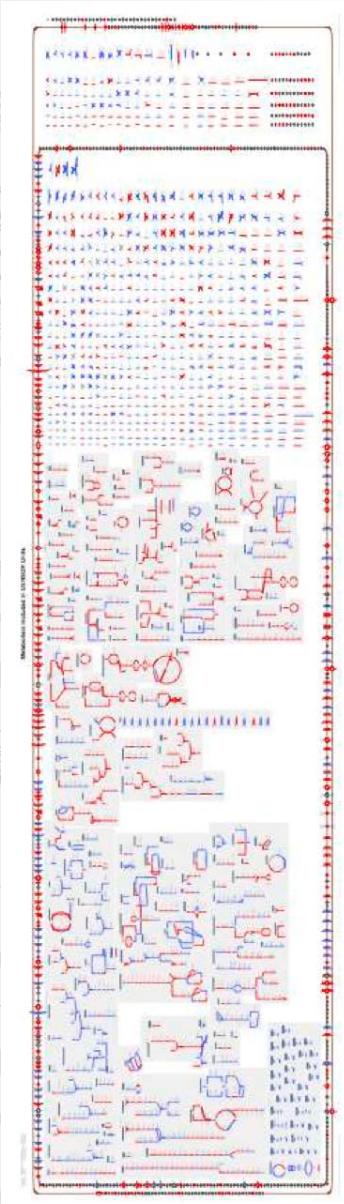
The TrpR GENSOR Unit



In the presence of L-tryptophan, TrpR represses transcription of genes that code for proteins necessary for synthesis and transport of tryptophan.

There is no known function for proteins AroM and YanjA.

Semi automatic construction of 189 GUs



frontiers

Genome-wide mapping of transcriptional regulation and metabolism describes information-processing units in Escherichia coli

Daniela Ledezma-Tejeida[1*], Cecilia Ishida[1], Julio Collado-Vides[1]

[1]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Mexico

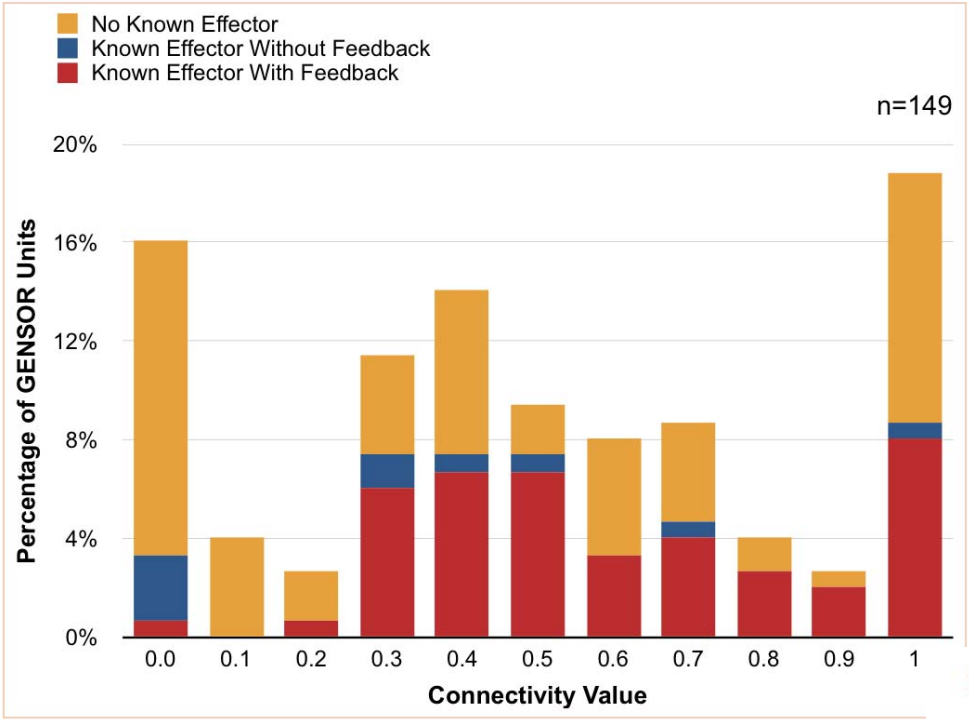
Submitted to Journal:
Frontiers in Microbiology

Specialty Section:
Systems Microbiology

Front Microbiol. 2017 Aug 3; 8: 1466.
doi: 10.3389/fmicb.2017.01466.

Table S3. Fraction of metabolism covered by GENSOR Units. Cellular Overview from Ecocyc showing in red the reactions included in the GENSOR Unit collection. Most metabolic pathways are covered, especially carbon, aminoacid and lipid metabolism. An SBML map of the 189 merged GENSOR Units is available on demand.

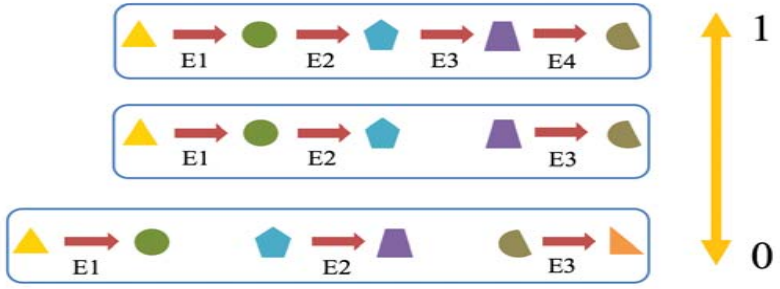
Connectivity of GUs response



Coneectivity

$$C = \frac{Ec}{Et + (Mft - 1)}$$

Where:
EC: connected enzymes
Et: Total number of enzymes
Mft: total number of metabolic fluxes

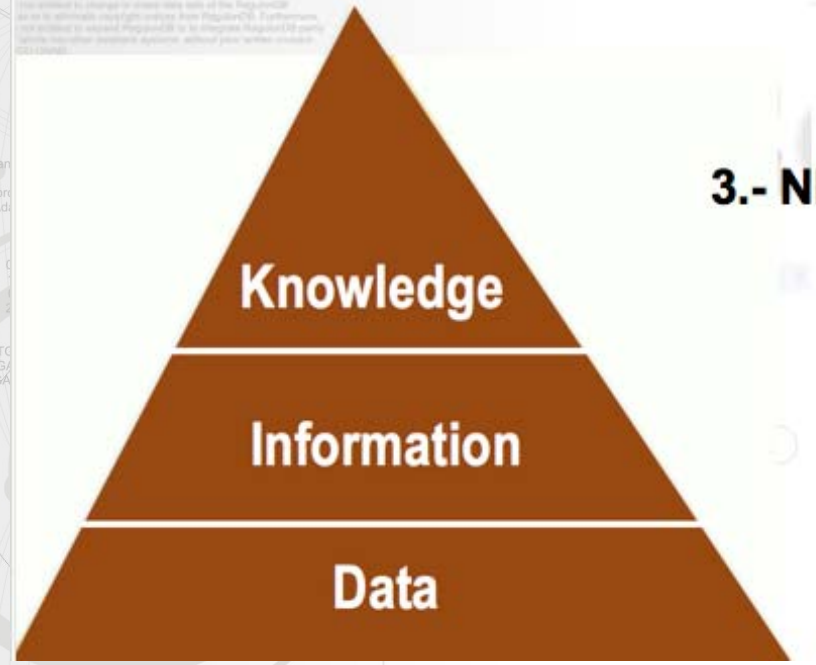


Extended GUs

GENSOR Unit	Data sets					
	RegulonDB	Galagan	ChIP-ChIP	ChIP-exo	SELEX	Version uploaded to RegulonDB
ArgR	0.64		0.64	0.64		(ChIP-exo) 0.34
CsiR	1.00	0.08				
Fur	0.52			0.36		(ChIP-exo) 0.51
LexA	0.40		0.24			
MntR	0.00				0.00	
Nac	0.45	0.42				
NsrR	0.43		0.06			
OmpR	0.11	0.07				(ChIP-exo) 0.46 (SELEX) 0.00
PurR	0.81		0.56			
RstA	0.00				0.20	
RutR	0.50		0.09			
UvrY	0.00			0.31		

Work in progress: many questions !

Natural language processing



3.- NLP strategies

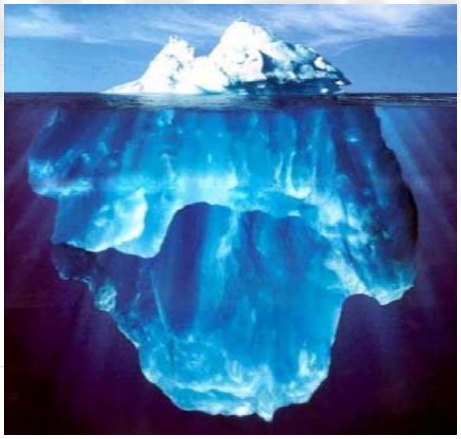
2.- Operons, regulons and GUs

1.- Curating HT experiments

The challenge: encoding "knowledge"

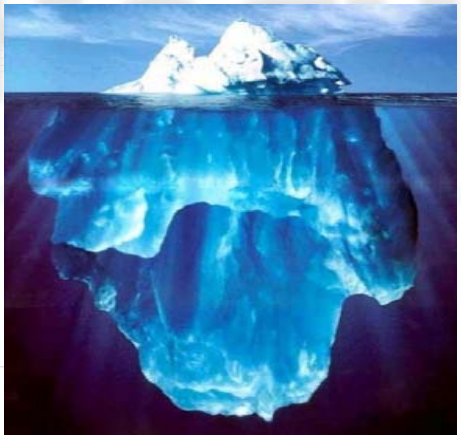
20 to 25%

75 to 80 %

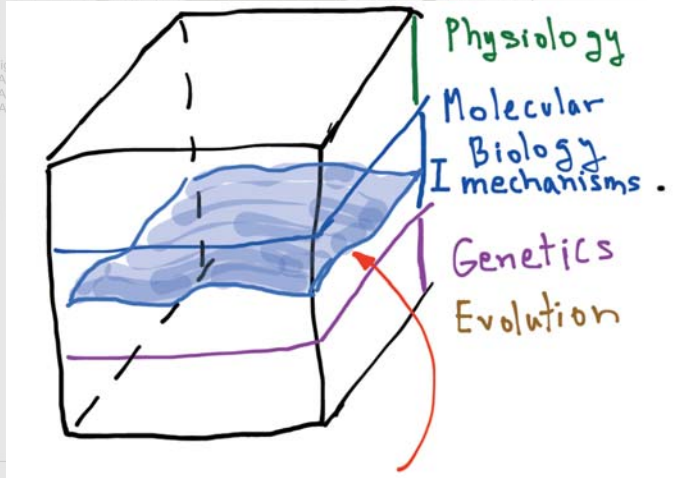


The challenge: encoding "knowledge"

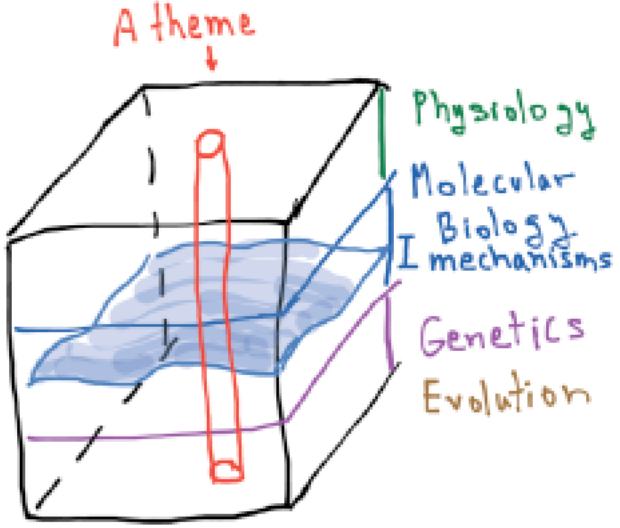
20 to 25%



75 to 80 %



we curate the shadows of mechanisms and transcription on the genome



How to enhance access to knowledge ?

We are using natural language processing tools

I. To enhance / improve curation

Curators

II. To enhance navigation through knowledge and its connection to data and information

Users

How to enhance access to knowledge ?

We are using natural language processing tools

Curators

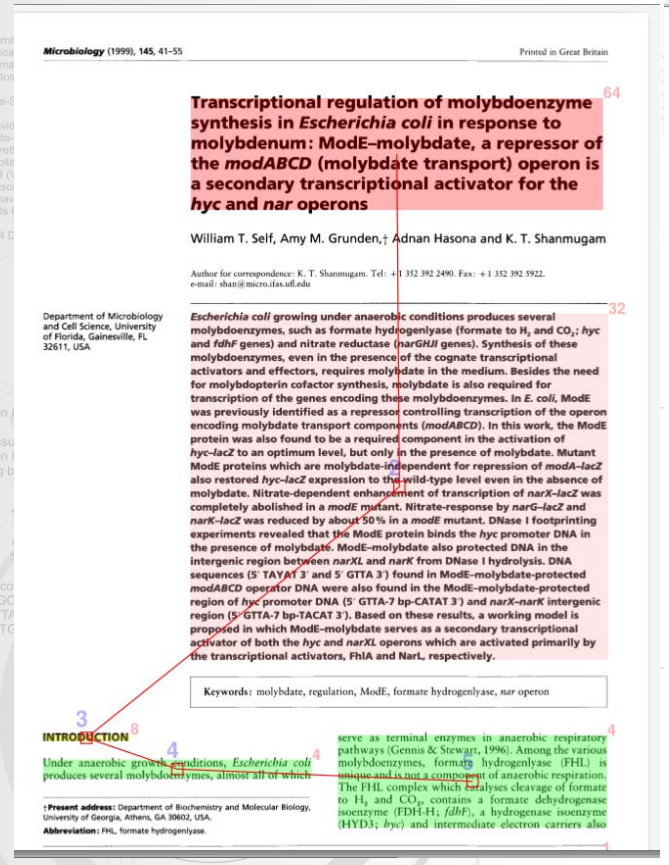
I. To enhance / improve curation

L-Regulon-DB

II. To enhance navigation through knowledge and its connection to data and information

Users

From PDF format to computer-accessible text files



Microbiology (1999), **145**, 41–55 Printed in Great Britain

Transcriptional regulation of molybdoenzyme synthesis in *Escherichia coli* in response to molybdenum: ModE–molybdate, a repressor of the *modABCD* (molybdate transport) operon is a secondary transcriptional activator for the *hyc* and *nar* operons

William T. Self, Amy M. Grunden,† Adnan Hasona and K. T. Shanmugam

Author for correspondence: K. T. Shanmugam. Tel: +1 352 392 2490. Fax: +1 352 392 5922. e-mail: shan@micro.ifas.ufl.edu

Department of Microbiology and Cell Science, University of Florida, Gainesville, FL 32611, USA

Escherichia coli **growing under anaerobic conditions produces several molybdoenzymes, such as formate hydrogenlyase (formate to H_2 and CO_2; *hyc* and *fdhF* genes) and nitrate reductase (*narGHJI* genes). Synthesis of these molybdoenzymes, even in the presence of the cognate transcriptional activators and effectors, requires molybdate in the medium. Besides the need for molybdopterin cofactor synthesis, molybdate is also required for transcription of the genes encoding these molybdoenzymes. In *E. coli*, ModE was previously identified as a repressor controlling transcription of the operon encoding molybdate transport components (*modABCD*). In this work, the ModE protein was also found to be a required component in the activation of *hyc–lacZ* to an optimum level, but only in the presence of molybdate. Mutant ModE proteins which are molybdate-independent for repression of *modA–lacZ* also restored *hyc–lacZ* expression to the wild-type level even in the absence of molybdate. Nitrate-dependent enhancement of transcription of *narX–lacZ* was completely abolished in a *modE* mutant. Nitrate-response by *narG–lacZ* and *narK–lacZ* was reduced by about 50% in a *modE* mutant. DNase I footprinting experiments revealed that the ModE protein binds the *hyc* promoter DNA in the presence of molybdate. ModE–molybdate also protected DNA in the intergenic region between *narXL* and *narK* from DNase I hydrolysis. DNA sequences (5′ TAYA1 3′ and 5′ GTTA 3′) found in ModE-molybdate-protected *modABCD* operator DNA were also found in the ModE–molybdate-protected region of *hyc* promoter DNA (5′ GTTA-7 bp-CATAT 3′) and *narX–narK* intergenic region (5′ GTTA-7 bp-TACAT 3′). Based on these results, a working model is proposed in which ModE–molybdate serves as a secondary transcriptional activator of both the *hyc* and *narXL* operons which are activated primarily by the transcriptional activators, FhlA and NarL, respectively.**

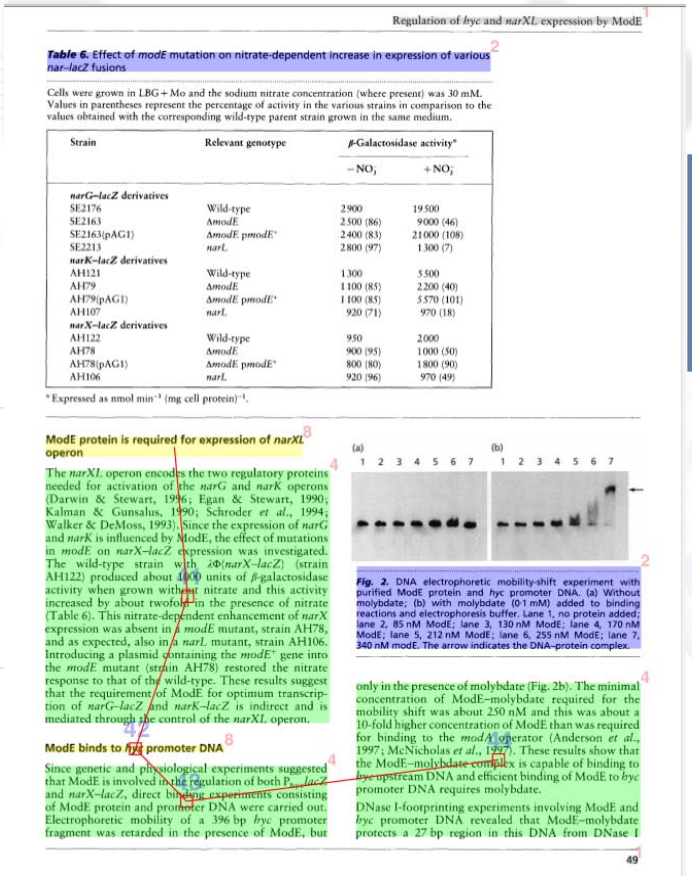
Keywords: molybdate, regulation, ModE, formate hydrogenlyase, *nar* operon

INTRODUCTION

Under anaerobic growth conditions, *Escherichia coli* produces several molybdoenzymes, almost all of which serve as terminal enzymes in anaerobic respiratory pathways (Gennis & Stewart, 1996). Among the various molybdoenzymes, formate hydrogenlyase (FHL) is unique and is not a component of anaerobic respiration. The FHL complex which catalyses cleavage of formate to H_2 and CO_2, contains a formate dehydrogenase isoenzyme (FDH-H; *fdhF*), a hydrogenase isoenzyme (HYD3; *hyc*) and intermediate electron carriers also

†**Present address:** Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602, USA.

Abbreviation: FHL, formate hydrogenlyase.



Regulation of *hyc* and *narXL* expression by ModE

Table 6. Effect of *modE* mutation on nitrate-dependent increase in expression of various *nar–lacZ* fusions

Cells were grown in LBG + Mo and the sodium nitrate concentration (where present) was 30 mM. Values in parentheses represent the percentage of activity in the various strains in comparison to the values obtained with the corresponding wild-type parent strain grown in the same medium.

Strain	Relevant genotype	β-Galactosidase activity*	
		$-NO_3^-$	$+NO_3^-$
***narG–lacZ* derivatives**			
SE2176	Wild-type	2900	19500
SE2163	Δ*modE*	2500 (86)	9000 (46)
SE2163(pAG1)	Δ*modE* p*modE*⁺	2400 (83)	21000 (108)
SE2213	*narL*	2800 (97)	1300 (7)
***narK–lacZ* derivatives**			
AH121	Wild-type	1300	5500
AH79	Δ*modE*	1100 (85)	2200 (40)
AH79(pAG1)	Δ*modE* p*modE*⁺	1100 (85)	5570 (101)
AH107	*narL*	920 (71)	970 (18)
***narX–lacZ* derivatives**			
AH122	Wild-type	950	2000
AH78	Δ*modE*	900 (95)	1000 (50)
AH78(pAG1)	Δ*modE* p*modE*⁺	800 (80)	1800 (90)
AH106	*narL*	920 (96)	970 (49)

*Expressed as nmol min^{-1} (mg cell protein)$^{-1}$.

ModE protein is required for expression of *narXL* operon

The *narXL* operon encodes the two regulatory proteins needed for activation of the *narG* and *narK* operons (Darwin & Stewart, 1996; Egan & Stewart, 1990; Kalman & Gunsalus, 1990; Schroder *et al.*, 1994; Walker & DeMoss, 1993). Since the expression of *narG* and *narK* is influenced by ModE, the effect of mutations in *modE* on *narX–lacZ* expression was investigated. The wild-type strain with λΦ(*narX–lacZ*) (strain AH122) produced about 1000 units of β-galactosidase activity when grown without nitrate and this activity increased by about twofold in the presence of nitrate (Table 6). This nitrate-dependent enhancement of *narX* expression was absent in a *modE* mutant, strain AH78, and as expected, also in a *narL* mutant, strain AH106. Introducing a plasmid containing the *modE*⁺ gene into the *modE* mutant (strain AH78) restored the nitrate response to that of the wild-type. These results suggest that the requirement of ModE for optimum transcription of *narG–lacZ* and *narK–lacZ* is indirect and is mediated through the control of the *narXL* operon.

ModE binds to *hyc* promoter DNA

Since genetic and physiological experiments suggested that ModE is involved in the regulation of both P_{hyc}–*lacZ* and *narX–lacZ*, direct binding experiments consisting of ModE protein and promoter DNA were carried out. Electrophoretic mobility of a 396 bp *hyc* promoter fragment was retarded in the presence of ModE, but

Fig. 2. DNA electrophoretic mobility-shift experiment with purified ModE protein and *hyc* promoter DNA. (a) Without molybdate; (b) with molybdate (0·1 mM) added to binding reactions and electrophoresis buffer. Lane 1, no protein added; lane 2, 85 nM ModE; lane 3, 130 nM ModE; lane 4, 170 nM ModE; lane 5, 212 nM ModE; lane 6, 255 nM ModE; lane 7, 340 nM ModE. The arrow indicates the DNA–protein complex.

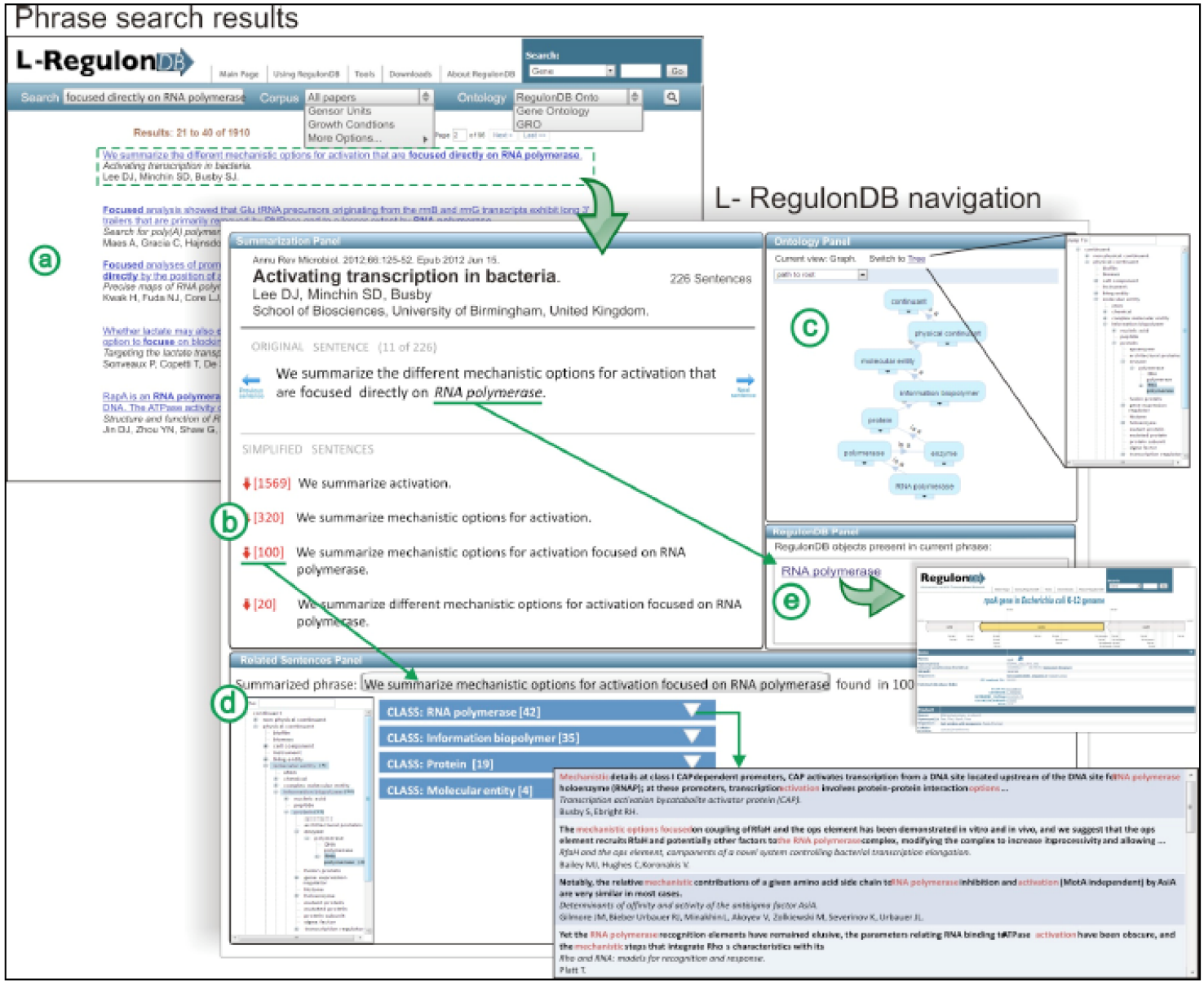
only in the presence of molybdate (Fig. 2b). The minimal concentration of ModE–molybdate required for the mobility shift was about 250 nM and this was about a 10-fold higher concentration of ModE than was required for binding to the *modA* operator (Anderson *et al.*, 1997; McNicholas *et al.*, 1997). These results show that the ModE–molybdate complex is capable of binding to *hyc* upstream DNA and efficient binding of ModE to *hyc* promoter DNA requires molybdate.

DNase I-footprinting experiments involving ModE and *hyc* promoter DNA revealed that ModE–molybdate protects a 27 bp region in this DNA from DNase I

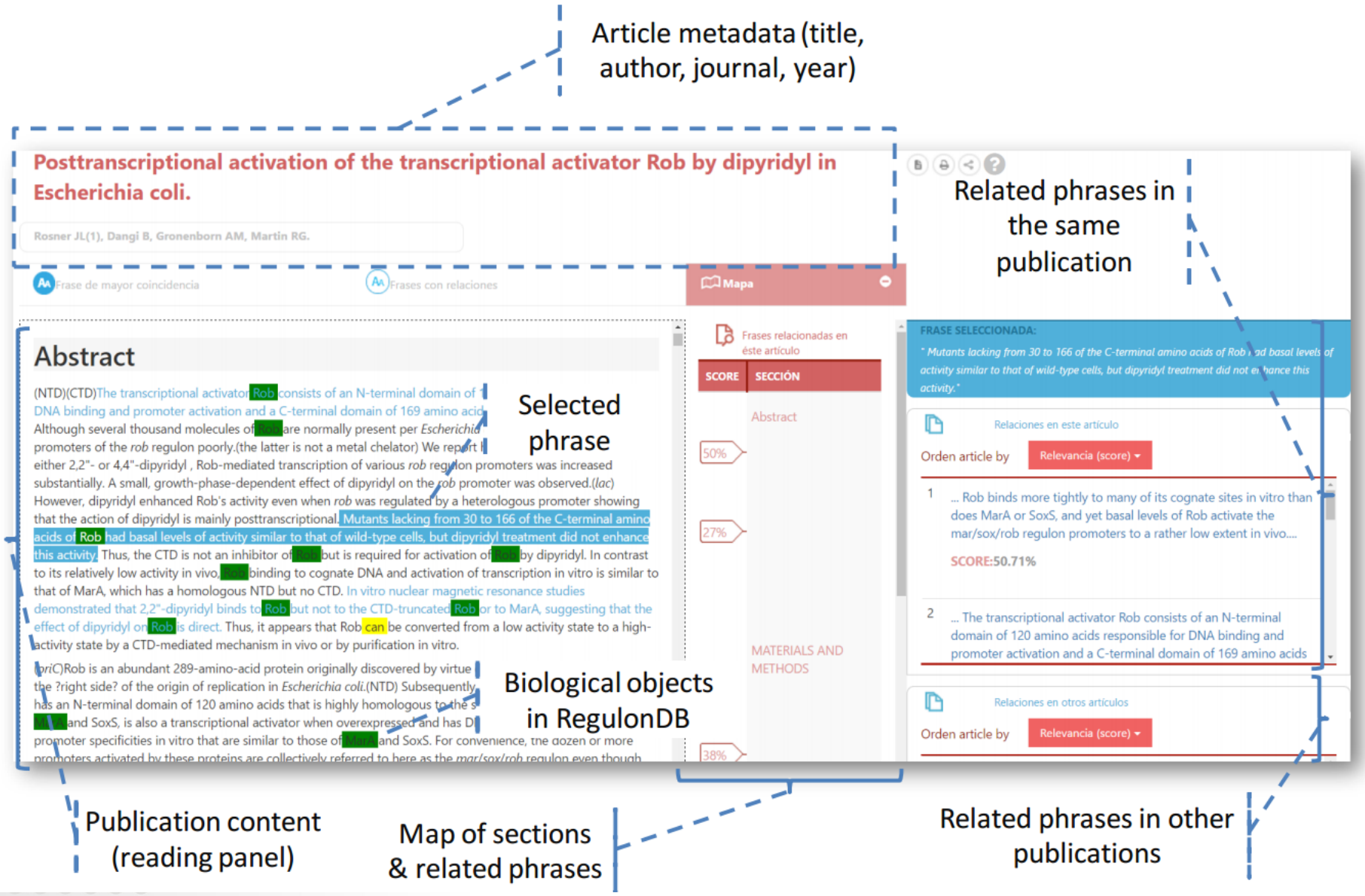
49

- Block identification
- Logical sentence extraction
- Reading order discovering

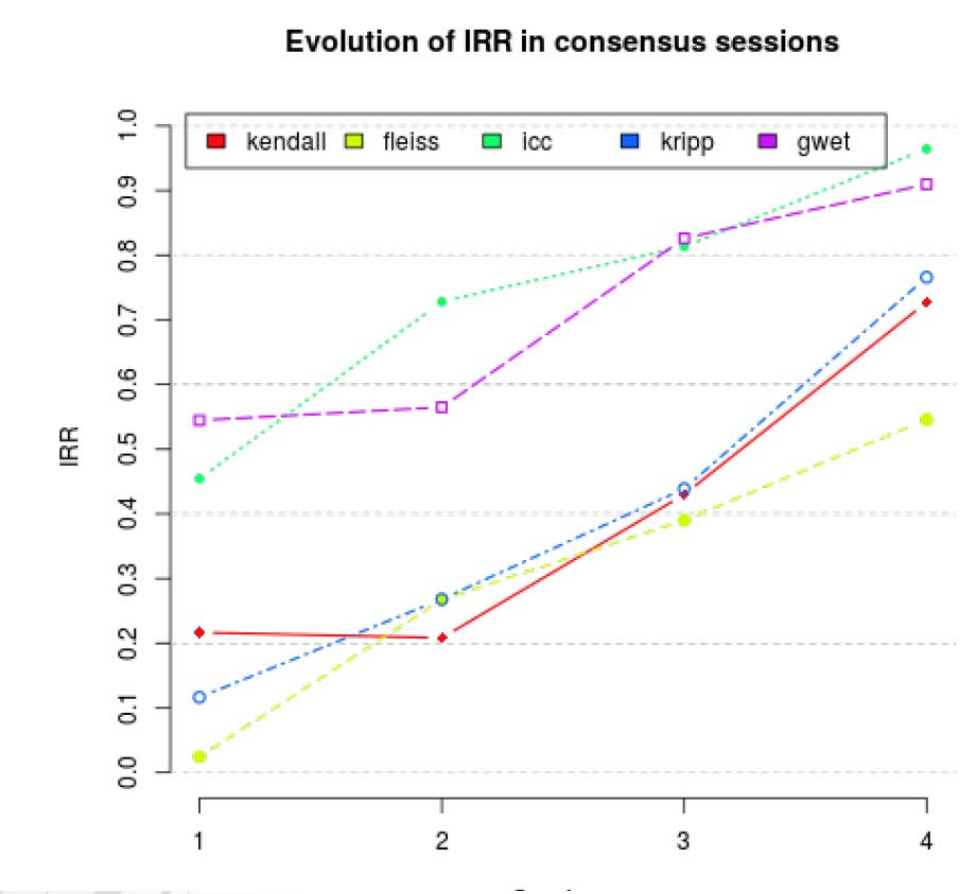
L-RegulonDB: a corpus for navigation



L-Regulon-DB
(Oscar Lithgow, Carlos Méndez, Fabio Rinaldi, Hilda Solano)



Similarity corpus on microbial transcriptional regulation



Inter-rating reliability (IRR) through agreement sessions

Lithgow-Serrano O.L., et al (2017) bioRxiv doi: 10.1101/219014

Three young friends

Xochicalco, 1994

Citation

User is committed to cite properly the use of RegulonDB. The current most update publication concerning RegulonDB is:
Socorro Gama-Castro, Verónica Jiménez-Jacinto, Martin Peralta-Gil, Alberto Santos-Zavaleta, Monica I. Peñaloza-Spindola, Bruno Contreras-Moreira,
Juan Segura-Salazar, Luis Muñiz-Rascado, Irma Martínez-Flores, Heladia Salgado,
César Bonavides-Martínez, Cei Abreu-Goodger, Carlos Rodriguez-Penagos, Juan Miranda-Rios,
Enrique Morett, Enrique Merino, Araceli M. Huerta, Luis Treviño-Quintanilla and Julio Collado-Vides.
"RegulonDB (Version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation"
Nucleic Acids Research, 2008, vol 36. D120-D124

Release: 6.4 Date: 10-AUG-09

Transcription Factor Matrix and Alignments

The consensus and patser programs were used to create the matrix and alignment
Transcription Factor NameAda
Total of uniq binding sites3

Matrix

A	2	3	1	1	3	2	3	0	0	1	1	2	0	0	1	0	2	2	0	3	2	0	1	1	0	0	0	1	0
C	1	0	0	1	0	1	0	0	1	0	2	0	0	2	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0
G	0	0	1	0	0	0	0	1	0	2	0	0	3	1	2	0	0	0	3	0	0	0	2	0	0	2	1	0	1
T	0	0	1	1	0	0	0	2	2	0	0	1	0	0	0	2	0	1	0	0	1	3	0	2	2	1	2	2	2

AlignmentScore
AAGCAAAGCGCAGCGTCTGAATAACGTTT20.66
AAAAAAATTAAAGCGCAAGATTGTTGGTT21.42
CATTACATTGCTGGATAAGAATGTTTTAG19.78

Xochicalco, 1994

And an old friend with great vision ! Congratulations Edgar !!

Acknowledgements

To all in the lab

Thank you for your remote attention

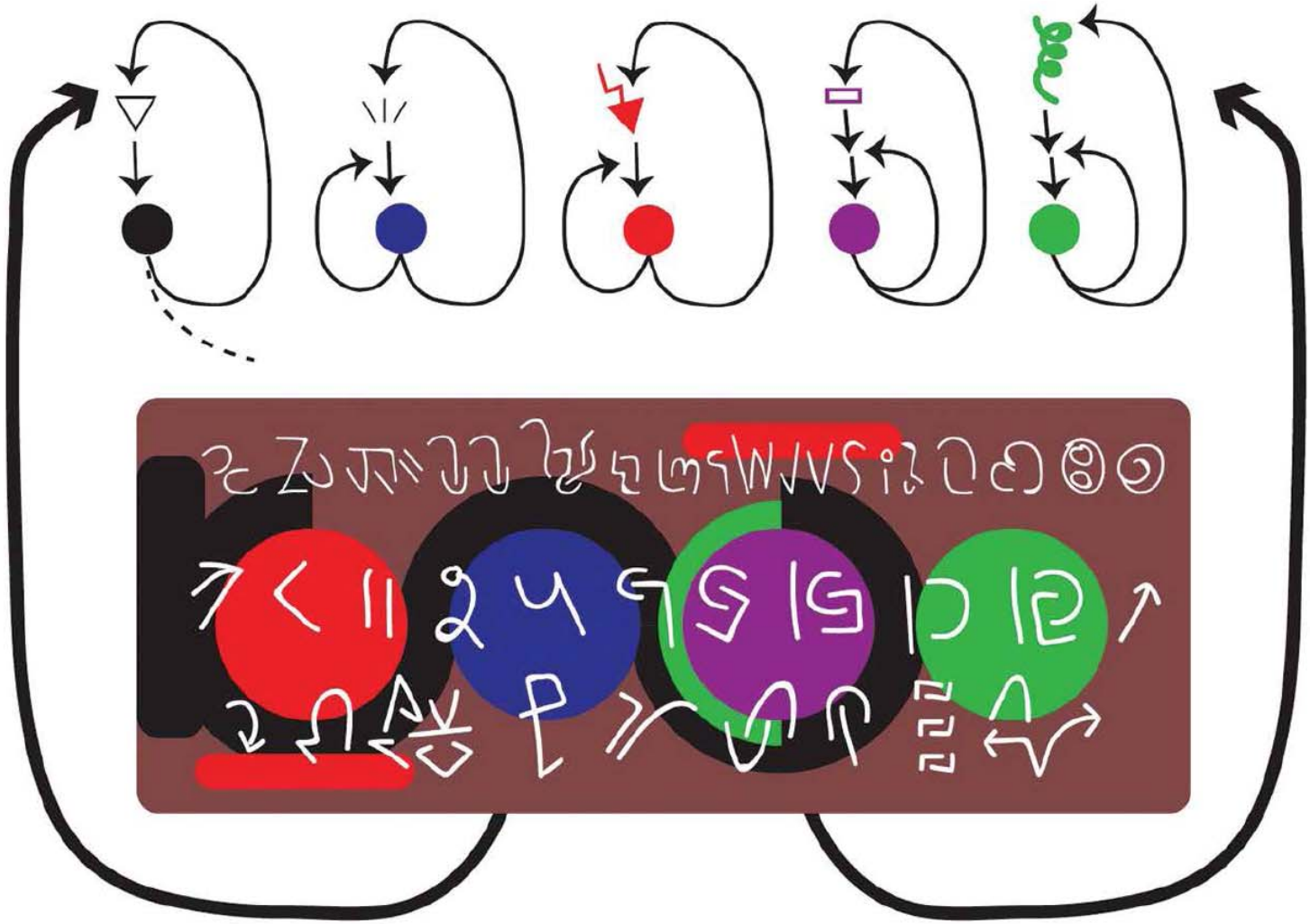
Funding supported by:

CONACYT
Consejo Nacional de Ciencia y Tecnología

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO
POR MI RAZA HABLARA EL ESPIRITU

NATIONAL INSTITUTES OF HEALTH

Codes, sensing and response



PGC- dic 2016