

*Regulatory Sequence Analysis*

# ***Transcription factor databases***

*Jacques.van.Helden@ulb.ac.be*

*Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)  
Université Libre de Bruxelles, Belgique*

# RegulonDB

## Transcriptional regulation in *Escherichia coli*

- RegulonDB Web site
  - <http://regulondb.ccg.unam.mx/>
- Model organism: *Escherichia coli*
- Data content
  - Transcription factors
  - Transcription factor binding sites (TFBS)
  - Position-specific scoring matrices (PSSM)
  - Promoters
  - Operons
- Collaboration with EcoCyc
  - EcoCyc is the reference database about metabolism in *Escherichia coli*
  - RegulonDB is integrated in the EcoCyc database



# Example of regulon in RegulonDB

RegulonDB 6.1

http://regulondb.ccg.unam.mx/RegulonControllerServlet?regulon\_id=ECK12R054548

**RegulonDB**  
Escherichia coli K12 Transcriptional Network

Main Page Using RegulonDB Tools Downloads About RegulonDB

Search: Gene  Go

Export to: XML

## REGULON LexA, in Escherichia coli K12 genome

**TRANSCRIPTION FACTOR:** [info](#)

**Name:** LexA [Tractordb tool](#)

**Connectivity:** Local Regulator

**Sensing Class:**

**Synonym(s):** LexA transcriptional repressor

**Gene name(s):** [lexA](#)

**Conformation(s):** LexA

**Coregulator(s):** [ArcA](#), [LexA](#)

**Note(s):** [Note\(s\): ...\[more\]](#)

**REGULATION EXERTED BY LexA,** [info](#)

Transcription Factor		Regulated		Binding Sites				Evidences	References
Conformation	Function	Promoter	Gene(s)	LeftPos	RigthPos	CenterPos	Sequence		
<b>KNOWN BINDING SITES (The centerpos is relative to the promoter +1)</b>									
<b>LexA repressor</b>									
LexA	repressor	lexAp	lexA,dinF	4255050	4255069	-50.5	ttcgataaatCTCTGGTTTATTGTGCAGTTtatggtcca	[HIBSCS]	[1] [2]
LexA	repressor	lexAp	lexA,dinF	4255091	4255110	-9	aatgccttTGCTGTATATACTACAGCAtaactgtata	[BPP] [HIBSCS]	[3] [2]
LexA	repressor	lexAp	lexA,dinF	4255112	4255131	13	ctcagcagcatAACTGTATATACACCCAGGggcggaatga	[BPP] [HIBSCS]	[3] [2]
LexA	repressor	phrBp	phr	738568	738587	-66.5	gccagcagctGGCTGCGCTTATCGACAGTTatgcctggg	[HIBSCS]	[1] [4]
LexA	repressor	phrBp	phr	738659	738678	25.5	ttatcctgacGCCTGGCTTTCAGGCGAGCgttatctgaa	[GEA] [HIBSCS]	[5]
LexA	repressor	rpsUp3	dnaG,rpoD,rpsU	3208751	3208770	4.5	attttgaaatAAGCTGGCGTTGATGCCAGCggcaaaccca	[BCE] [SM]	[6] [7]
LexA	repressor	recAp	recA,recX	2821851	2821870	-21	aaacactgaTACTGTATGAGCATACAGTAtaattgctc	[BPP]	[3] [2]
LexA	repressor	sulAp	sulA	1020162	1020181	-2	ctggatgtacTGTACATCCATACAGTAACTcacaggggct	[BCE]	[8]
LexA	repressor	uvrBp2	uvrB	812655	812674	-20	tatggtgatgAACTGTTTTTTATCCAGTAtaattgttg	[HIBSCS]	[9]
LexA	repressor	uvrDp1	uvrD	3995930	3995949	11	taatcagcaaATCTGTATATATACCCAGCTtttggcgga	[HIBSCS]	[10]
LexA	repressor	umuDp	umuC,umuD	1229951	1229970	-0.5	aagaacagacTACTGTATATAAAAACAGTAtaacttcagg	[BPP]	[11] [1]
LexA	repressor	umuDp	umuC,umuD	1229931	1229950	-20.5	atcaglatgATCTGCTGGCAAGAACAGACtactgtatat	[HIBSCS]	[11] [1] [12]
LexA	repressor	insKp	insK					[AIBSCS] [GEA]	[13]
LexA	repressor	dinQp	dinQ					[AIBSCS] [GEA]	[13]
LexA	repressor	polBp	polB	65834	65853	-40.5	gggcagtaatGACTGTATAAAACCACAGCCaatcaaacga	[AIBSCS] [GEA]	[14]
LexA	repressor	ruvAp1	ruvA,ruvB	1944102	1944121	-63.5	aataaattACTGTGCCATTTTTCAGTTcatcgagacac	[HIBSCS]	[15] [1]
LexA	repressor	ruvAp1	ruvA,ruvB	1944051	1944070	-12.5	tctcatcctTCGCTGGATATCTATCCAGCattttttat	[BPP] [HIBSCS]	[15] [13] [1] [16]
LexA	repressor	ruvAp2	ruvA,ruvB	1944103	1944122	-72.5	gaataaattaTACTGTGCCATTTTTCAGTTcatcgagaca	[HIBSCS]	[15] [1]

Done

# PSSM in RegulonDB

- RegulonDB contains a collection of PSSM built by aligning annotated binding sites.
  - [http://regulondb.ccg.unam.mx/data/Matrix\\_AlignmentSet.txt](http://regulondb.ccg.unam.mx/data/Matrix_AlignmentSet.txt)
- This collection can be used to scan genomes and predict new TFBS.

```
...
-----

Transcription Factor Name    LexA
Total of uniq binding sites  23

Matrix
A   12   0   0   0   1  12   1  12   6  10   7  13   4  12   0  23   0   1  12   6  11
C    3  22   0   0   2   3   5   2   2   5   5   2   4   7  23   0   0   8   2   2   3
G    5   0   0  23   6   3   2   4   0   2   0   3   3   2   0   0  23   1   3   2   1
T    3   1  23   0  14   5  15   5  15   6  11   5  12   2   0   0   0  13   6  13   8

Alignment      Score
ACTGTATAAAACACAGCCAA      12.05
GCTGCGCTTATCGACAGTTAT      8.48
CCTGGCTTTCAGGGCAGCGTT      7.51
ACTGTTTTTTTATCCAGTATA      16.18
ATTGGCTGTTTATACAGTATT      12.01
CCTGTTAATCCATACAGCAAC      10.7
ACTGTACATCCATACAGTAAC      14.66
TCTGCTGGCAAGAACAGACTA      3.36
ACTGTATATAAAAACAGTATA      17.23
GCTGGATATCTATCCAGCATT      15.55
GCTGGATATCTATCCAGCATT      15.55
ACTGTGCCATTTTTCAGTTCA      8.61
ACTGTGCCATTTTTCAGTTCA      8.61
ACTGTATATAAAACAGTTTA      16.16
ACTGTACACAATAACAGTAAT      12.47
ACTGTATGAGCATACAGTATA      14.73
GCTGGCGTTGATGCCAGCGGC      4.27
ACTGTTTATTTATACAGTAAA      16.67
TCTGTATATATACCCAGCTTT      14.73
TCTGGTTTATTGTGCAGTTTA      9.97
GCTGTATATACTCACAGCATA      15.05
ACTGTATATACACCCAGGGGG      9.28
CCTGAATGAATATACAGTATT      12.9

-----
....
```

# TRANSFAC - Gene transcription factor database

- Organisms
  - Eukaryotes
  - Particular emphasis on mammals (specially human, mouse, rat)
- Distribution
  - The public version is not updated anymore
  - Commercial version (TRANSFAC PRO)
  - Distributed by BioBase™
    - <http://www.biobase.de/>
- Data content
  - Transcription factors
  - Binding sites
    - Evidences
    - Publications
  - Position-specific scoring matrices
  - Pattern matching tools (*patch*, *match*)

BIOBASE Biological Databases: TRANSFAC Gene Transcription Factor Database

<http://www.biobase.de/pages/index.php?id=transfac>

**BIOBASE**  
BIOLOGICAL DATABASES

BIOBASE Knowledge Library

Home • Contact Form • Sitemap • Imprint • Privacy Policy • Free Trials • Login

You are here : BIOBASE Knowledge Library / TRANSFAC databases / TRANSFAC

- > Company
- > BIOBASE Knowledge Library
  - > BKL PLANT Edition
  - > TRANSFAC databases
    - > TRANSFAC
      - Free Trial
      - Statistics
      - Brochure
      - Proof of Principle
      - Guided Tour
      - References
      - Information Request
- > TRANSCompel
- > TRANSPro
- > PathoDB
- > S/MARt DB
- > TRANSPATH database
- > PROTEOME databases
- > BRENDA Professional
- > HGMD Professional
- > Applications
- > Custom Services
- > Subscriptions
- > Support
- > News
- > Contact

Search

## TRANSFAC Gene Transcription Factor Database

**TRANSFAC® - Gene Transcription Factor Database**

TRANSFAC - the internationally unique knowledge base - contains data on transcription factors, their experimentally-proven binding sites, and regulated genes. Its broad compilation of binding sites allows the derivation of positional weight matrices.

TRANSFAC's programs, *Match* and *Patch*, use the matrices and the site sequences themselves for performing the matrix-or pattern-based search of factor binding sites in regulatory DNA sequences. Thus, it is possible to make predictions for most gene promoters, which have not been studied in detail yet.

TRANSFAC also includes a tool to automatically visualize gene-regulatory networks being based on interlinked factor and gene entries in the database (gene regulation and gene expression).

In addition, TRANSFAC comprises

- extensive information on transcription factors and their structures, functions, expression patterns
- a recently added table for invivo binding sequences from ChIP on chip experiments.

**Affymetrix GeneChip® Compatibility**

TRANSFAC works in conjunction with our ExPlain™ analysis system to apply a new knowledge driven approach to the analysis of whole complexes of coexpressed genes. The internal Composite Module Analyst (CMA) is a genetic algorithm for analysis and prediction of relevant promoters in the identified set of given genes obtained for sources such as Affymetrix GeneChip® Arrays. This combinatorial analysis drops false positive rates significantly and enables scientists to find potential causes for specific cellular events.

The power of correct prediction in ExPlain is driven by TRANSFAC®, a knowledge base of high quality, expert level, manually curated published scientific literature. TRANSFAC presents data on transcription factors, their experimentally-proven binding sites, and regulated genes. To

Latest News

- April 14, 08  
Using BIOBASE tools to identify biomarkers

Upcoming Events

- May 19-21, 2008  
Biomarker World Congress

Loews Philadelphia Hotel - Philadelphia, PA

Join us...

# ORegAnno

- All organisms (with specific focus on metazoan)
- Web site
  - <http://www.oreganno.org/oreganno/Index.jsp>
  - Also available from the UCSC genome browser
    - <http://genome.ucsc.edu/>
- Community-based annotation (Jamboree)
- Data content
  - Transcription factor binding sites
  - Mapping on the genomes

The screenshot shows the ORegAnno website in a web browser. The browser's address bar displays <http://www.oreganno.org/oreganno/Index.jsp>. The website has a header with the ORegAnno logo and the tagline "open regulatory annotation database". A navigation menu on the right includes links for "login", "new user", "logout", "search", "annotate", "queue", "tools", "dump", "help", and "cite". The main content area is titled "AN OPEN ACCESS DATABASE FOR GENE REGULATORY ELEMENT AND POLYMORPHISM ANNOTATION" and contains a detailed description of the database, its purpose, and its data sources. It also includes a "NEWS" section with recent updates and a "MOST RECENTLY ANNOTATED PUBLICATIONS" section with a list of recent research papers. At the bottom, there is a section for "INCORPORATED DATASETS" and a footer with the text "ORegAnno provides the ability to incorporate well-established datasets and provide relevant citation and ongoing maintenance".

ORegAnno: Open Regulatory Annotation

<http://www.oreganno.org/oreganno/Index.jsp>

You are not logged in

REGULATORY HAPLOTYPE: 7 entries.  
REGULATORY REGION: 25994 entries.  
TRANSCRIPTION FACTOR BINDING SITE: 14475 entries.  
REGULATORY POLYMORPHISM: 175 entries.

More details...

menu

- login
- new user
- logout

user menu

- search
- annotate
- queue
- tools
- dump
- help
- cite

AN OPEN ACCESS DATABASE FOR GENE REGULATORY ELEMENT AND POLYMORPHISM ANNOTATION

The Open REGULATORY ANNOTATION database (ORegAnno) is an open database for the curation of known regulatory elements from scientific literature. Annotation is collected from users worldwide for various biological assays and is automatically cross-referenced against PubMed, Entrez Gene, Ensembl, dbSNP, the eVOC, Cell type ontology, and the Taxonomy database, where appropriate, with information regarding the original experimentation performed (evidence). ORegAnno further provides an open validation process for all regulatory annotation in the public domain. Assigned validators receive notification of new records in the database and are able to cross-reference the citation to ensure record integrity. Validators have the ability to modify any record (deprecating the old record and creating a new one) if an error is found. Further, any contributor to the database can comment on any annotation by marking errors, or adding special reports into function as they see fit. These features of ORegAnno ensure that the collection is of the highest quality and uniquely provides a dynamic view of our changing understanding of gene regulation in the various genomes. As a first step, we recommend reading through our Help page.

The ORegAnno data and web application are all LGPL open-source to encourage the development and maintenance of the database to new information and experimentation techniques. Please use our current citation information when referring to ORegAnno data in publication. We encourage interested contributors to send email to the ORegAnno mailing list at [oreganno@bcgsc.ca](mailto:oreganno@bcgsc.ca) or to visit the mailing-list archives.

NEWS

**October 17th, 2007** Warning: there are persistent case sensitivity issues with the Boolean search features. A fix will be available shortly.

**September 21st, 2007** There will be an interruption of ORegAnno service on the 21st as the BCGSC systems undergo maintenance

**July 26th, 2007** A set of NRSF/REST binding sites identified using ChIPSeq by Caltech/Stanford has been added [\[more\]](#)

[More news...](#)

MOST RECENTLY ANNOTATED PUBLICATIONS

Gao H et al., Genome-wide identification of estrogen receptor alpha-binding sites in mouse liver. *Mol Endocrinol* 2008

Harblson CT et al., Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004

Lim CA et al., Genome-wide mapping of RELA(p65) binding identifies E2F1 as a transcriptional activator recruited by NF-kappaB upon TLR4 activation. *Mol Cell* 2007

Lin CY et al., Whole-genome cartography of estrogen receptor alpha binding sites. *PLoS Genet* 2007

MacIsaac KD et al., An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* 2006

INCORPORATED DATASETS

ORegAnno provides the ability to incorporate well-established datasets and provide relevant citation and ongoing maintenance

Done

## *Other databases (to develop in further versions of this course)*

---

- YeasTract <http://www.yeasttract.com/>
  - Yeast-specific database. Factors, binding sites and motifs + tools.
- JASPAR <http://jaspar.cgb.ki.se/>
  - Essentially PSSMs for vertebrates
- FlyReg <http://www.flyreg.org/>
  - Drosophila DNase I Footprint Database
- PlantCARE <http://bioinformatics.psb.ugent.be/webtools/plantcare/html/>
  - Plant Cis-Acting Regulatory Elements