Regulatory Sequence Analysis

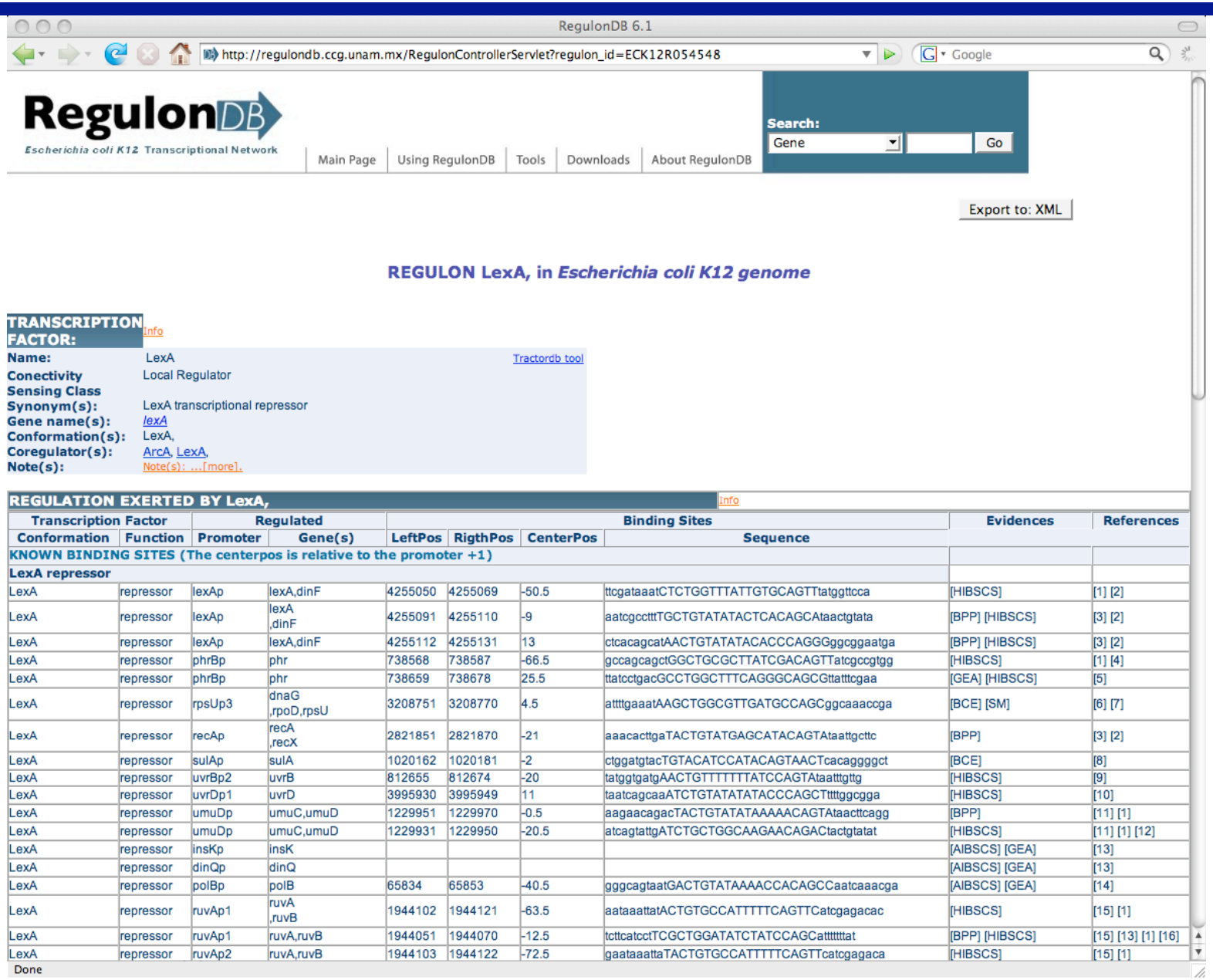
Transcription factor databases

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



Transcription Factor		Regulated		Binding Sites				Evidences	References
Conformation	Function	Promoter	Gene(s)	LeftPos	RigthPos	CenterPos	Sequence		
KNOWN BINDING SITES (The centerpos is relative to the promoter +1)									
LexA repressor									
LexA	repressor	lexAp	lexA,dinF	4255050	4255069	-50.5	ttcgataaatCTCTGGTTTATTGTGCAGTTtatggttcca	[HIBSCS]	[1] [2]
LexA	repressor	lexAp	lexA ,dinF	4255091	4255110	-9	aatcgcctttTGCTGTATATACTCACAGCAtaactgtata	[BPP] [HIBSCS]	[3] [2]
LexA	repressor	lexAp	lexA,dinF	4255112	4255131	13	ctcacagcatAACTGTATATACACCCAGGGggcggaatga	[BPP] [HIBSCS]	[3] [2]
LexA	repressor	phrBp	phr	738568	738587	-66.5	gccagcagctGGCTGCGCTTATCGACAGTTatcgccgtgg	[HIBSCS]	[1] [4]
LexA	repressor	phrBp	phr	738659	738678	25.5	ttatcctgacGCCTGGCTTTCAGGGCAGCGttatttcgaa	[GEA] [HIBSCS]	[5]
LexA	repressor	rpsUp3	dnaG ,rpoD,rpsU	3208751	3208770	4.5	attttgaaatAAGCTGGCGTTGATGCCAGCggcaaaccga	[BCE] [SM]	[6] [7]
LexA	repressor	recAp	recA ,recX	2821851	2821870	-21	aaacacttgaTACTGTATGAGCATACAGTAtaattgcttc	[BPP]	[3] [2]
LexA	repressor	sulAp	sulA	1020162	1020181	-2	ctggatgtacTGTACATCCATACAGTAACTcacaggggct	[BCE]	[8]
LexA	repressor	uvrBp2	uvrB	812655	812674	-20	tatggtgatgAACTGTTTTTTTATCCAGTAtaatttgttg	[HIBSCS]	[9]
LexA	repressor	uvrDp1	uvrD	3995930	3995949	11	taatcagcaaATCTGTATATATACCCAGCTttttggcgga	[HIBSCS]	[10]
LexA	repressor	umuDp	umuC,umuD	1229951	1229970	-0.5	aagaacagacTACTGTATATAAAAACAGTAtaacttcagg	[BPP]	[11] [1]
LexA	repressor	umuDp	umuC,umuD	1229931	1229950	-20.5	atcagtattgATCTGCTGGCAAGAACAGACtactgtatat	[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	aataaattatACTGTGCCATTTTTCAGTTCatcgagacac	[HIBSCS]	[15] [1]
LexA	repressor	ruvAp1	ruvA,ruvB	1944051	1944070	-12.5	tcttcatcctTCGCTGGATATCTATCCAGCatttttttat	[BPP] [HIBSCS]	[15] [13] [1] [16]
LexA	repressor	ruvAp2	ruvA,ruvB	1944103	1944122	-72.5	gaataaattaTACTGTGCCATTTTTCAGTTcatcgagaca	[HIBSCS]	[15] [1]

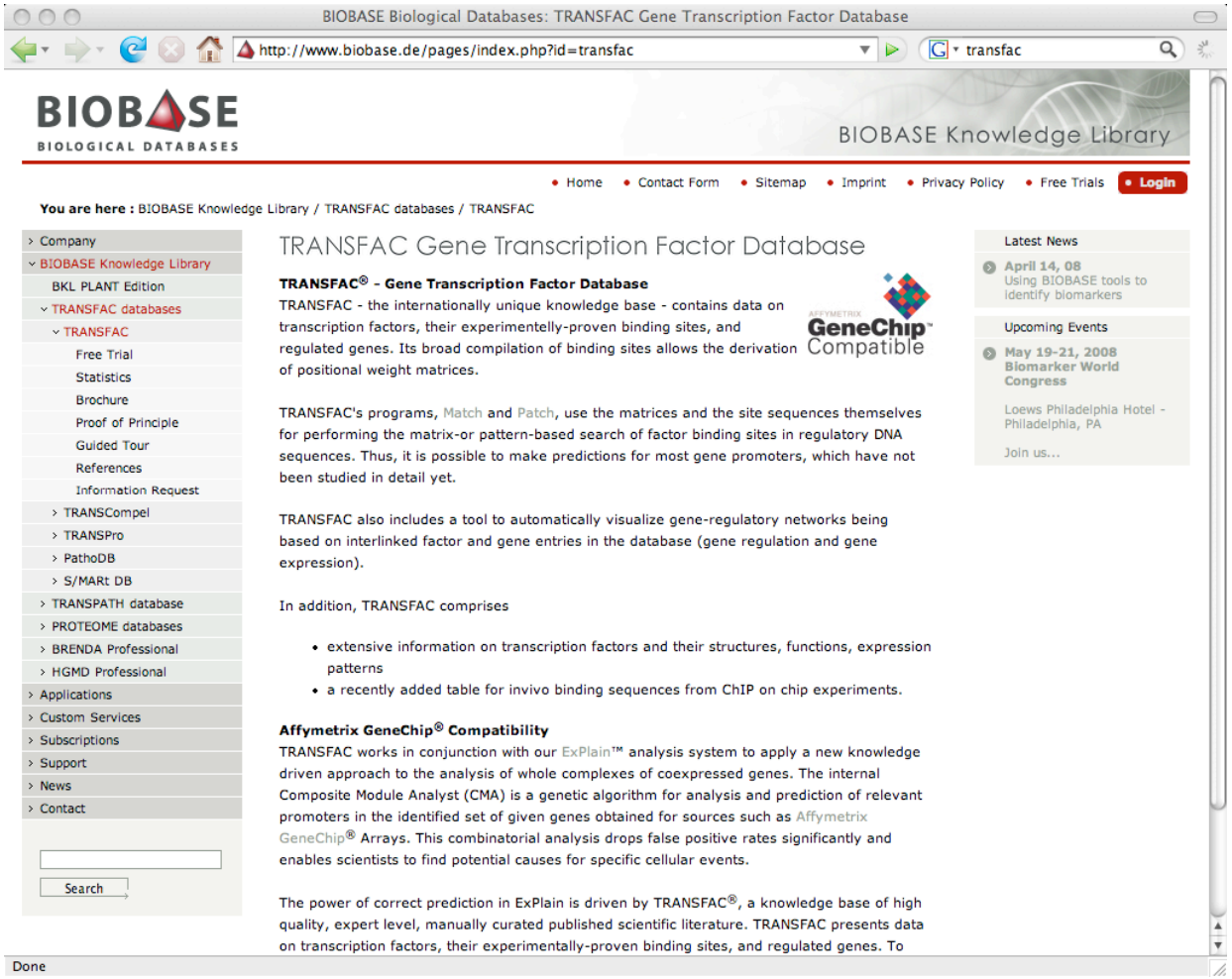
PSSM in RegulonDB

- RegulonDB contains a collection of PSSM built by aligning annotated binding sites.
  - 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
ACTGTATAAAACCACAGCCAA        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
ACTGTATATAAAACCAGTTTA        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

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

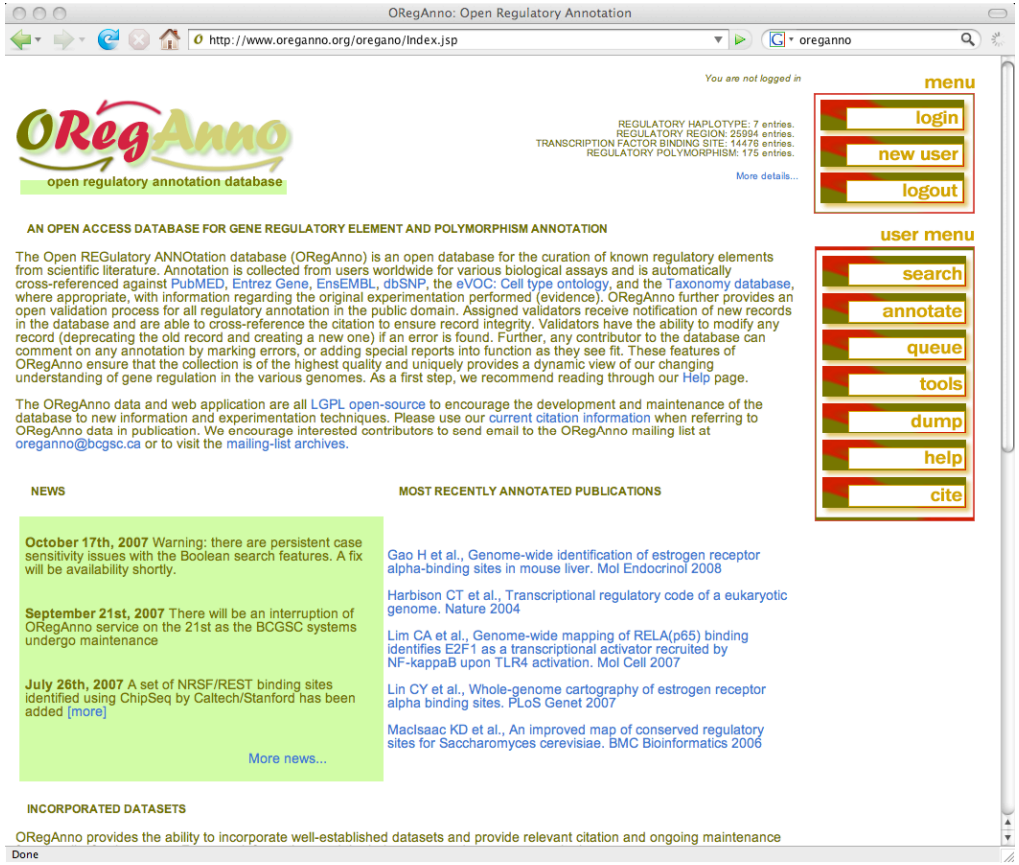
TRANSFAC - Gene transcription factor database

- 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*)



ORegAnno

- All organisms (with specific focus on metazoan)
- Web site
  - http://www.oreganno.org/oregano/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



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

- YeasTract http://www.yeastract.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