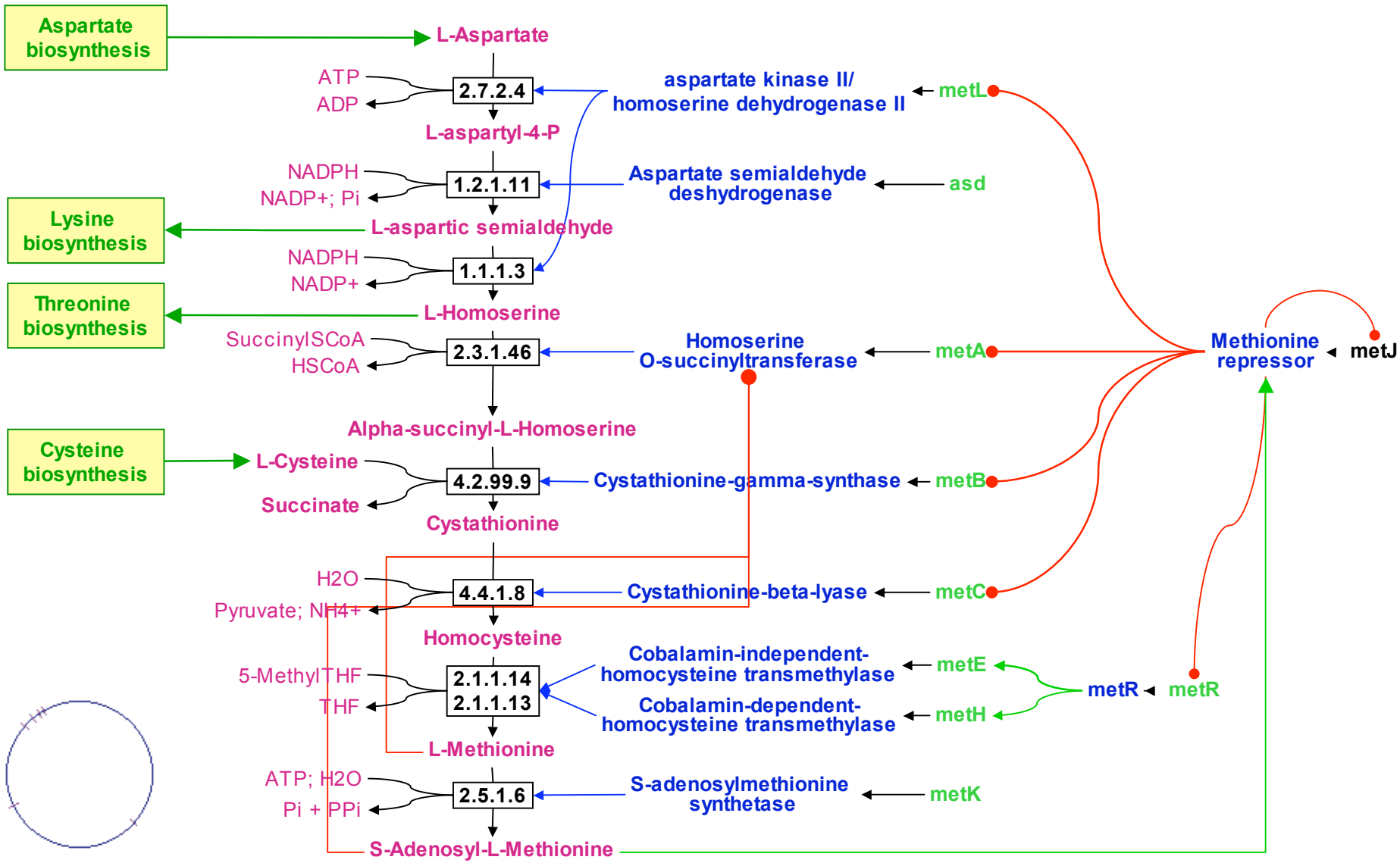
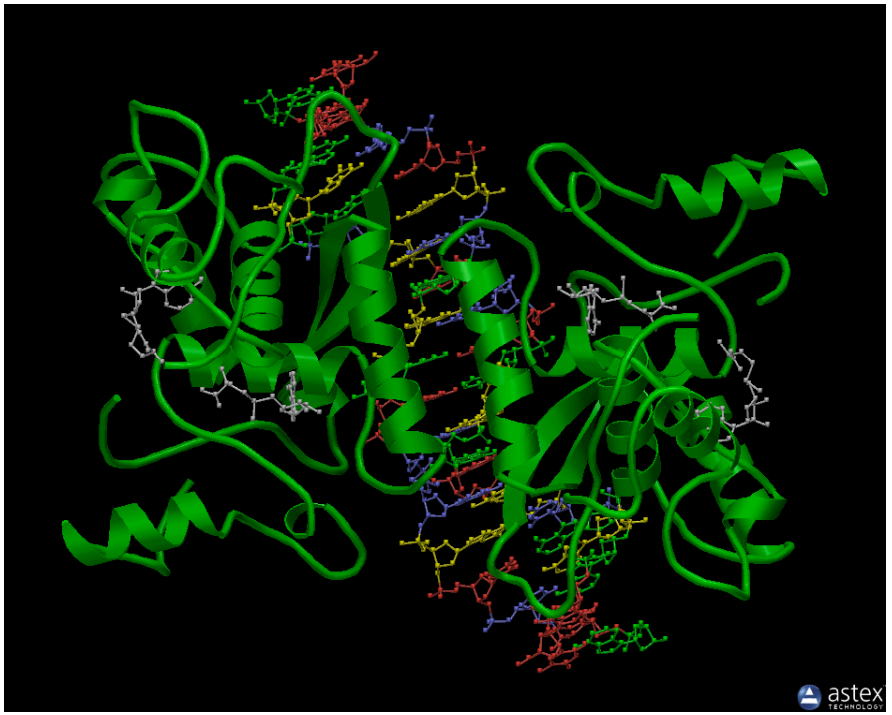
From phylogenetic footprints to co-regulation networks

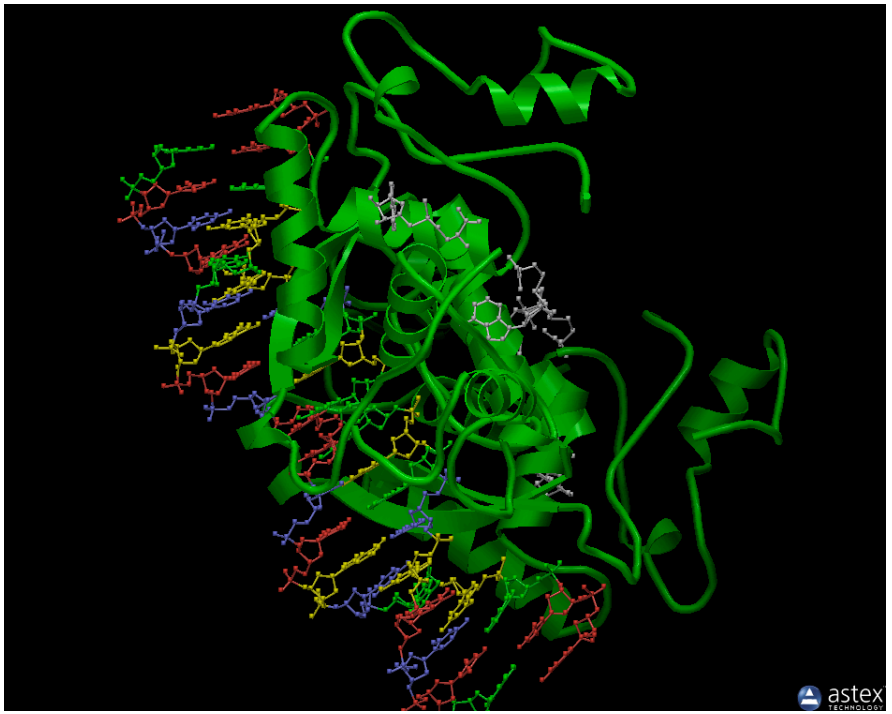
Methionine Biosynthesis in E.coli



Methionine repressor

- Structure cristalline du répresseur de la méthionine chez *Escherichia coli*.
- En vert: la protéine MetJ forme un dimère qui se lie à l'ADN
- La séquence d'ADN est colorée par nucléotide
- En gris: molécules de méthionine liées au répresseur (elles activent le répresseur)



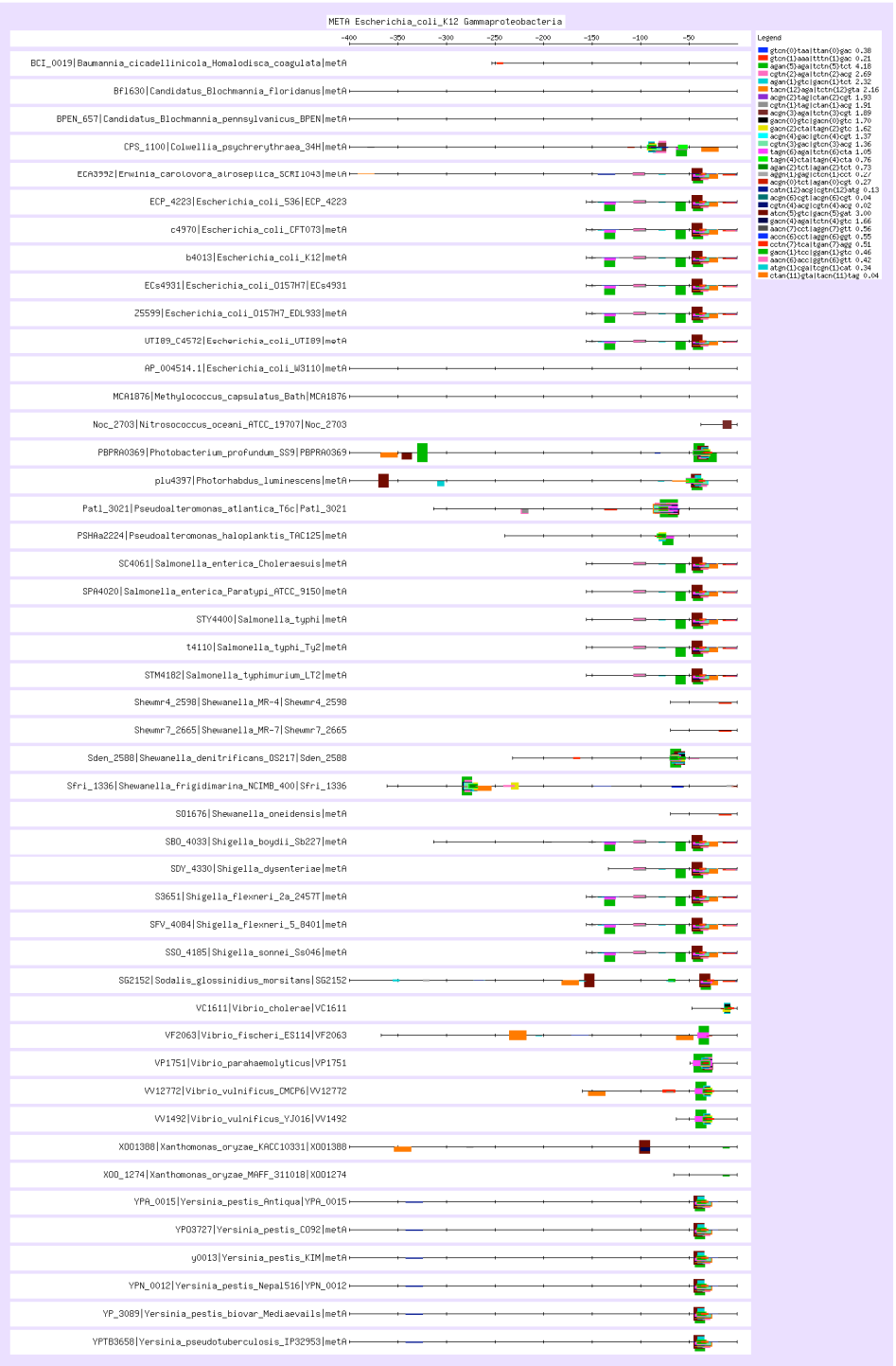


http://www.ebi.ac.uk/msd-srv/msdlite/apps/query?id_type=swissProtId&id_value=P0A8U6

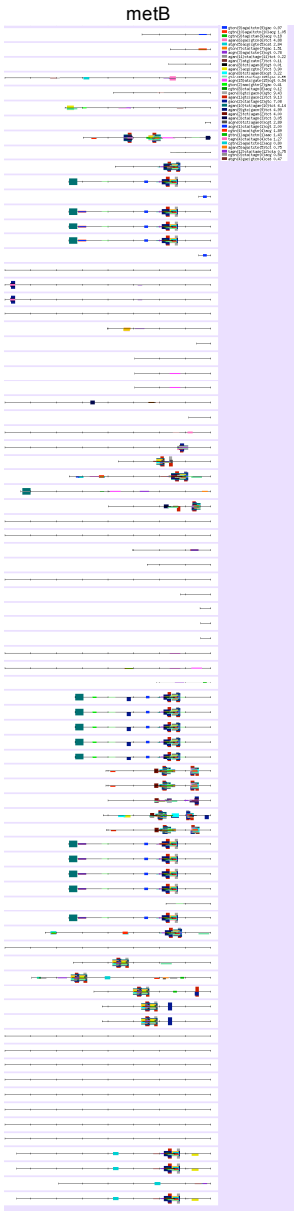
Phylogenetic footprints

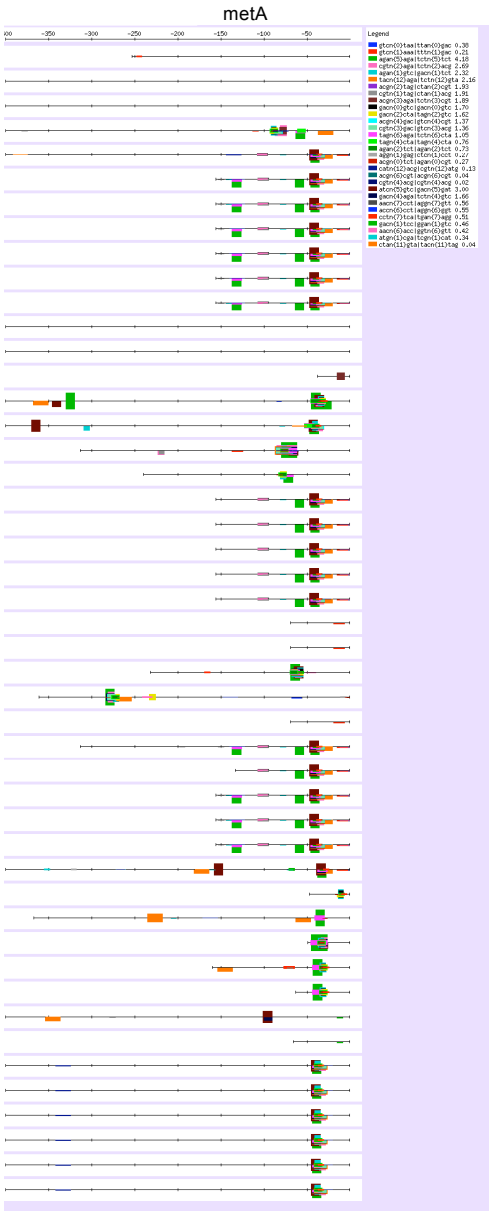
MetA orthologs in Gammaproteobacteria

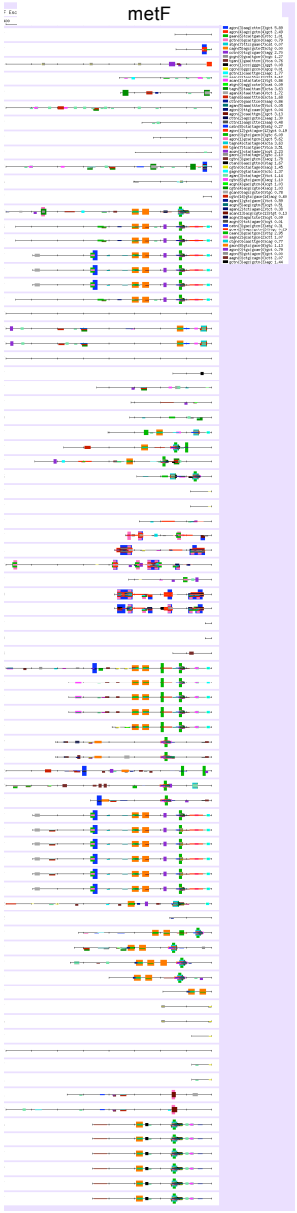
atcnnnnngtc.....	.....gacnnnnngat	3.00
.tctnnnnntct....	....agannnnnaga.	4.18
.tctnnacg.......	.......cgtnnaga.	2.69
.tctnnncgt......	......acgnnnaga.	1.89
.tctnnnngtc.....	.....gacnnnnaga.	1.66
.tctnnnnnncta...	...tagnnnnnnaga.	1.05
..ctanncgt......	......acgnntag..	1.93
..ctanacg.......	.......cgtntag..	1.91
...tagnngtc.....	.....gacnncta...	1.62
...tagnnnnnnaga.	.tctnnnnnncta...	1.05
...tagnnnncta...	...tagnnnncta...	0.76
....agannnnnaga.	.tctnnnnntct....	4.18
....agangtc.....	.....gacntct....	2.32
....aganntct....	....aganntct....	0.73
....agacgt......	......acgtct....	0.27
.....gacnnnnngat	atcnnnnngtc.....	3.00
.....gacntct....	....agangtc.....	2.32
.....gacgtc.....	.....gacgtc.....	1.70
.....gacnnnnaga.	.tctnnnngtc.....	1.66
.....gacnncta...	...tagnngtc.....	1.62
......acgnntag..	..ctanncgt......	1.93
......acgnnnaga.	.tctnnncgt......	1.89
......acgtct....	....agacgt......	0.27
.......cgtnnaga.	.tctnnacg.......	2.69
.......cgtntag..	..ctanacg.......	1.91
atctagacgtctagat	**atctagacgtctagat**	**4.18**

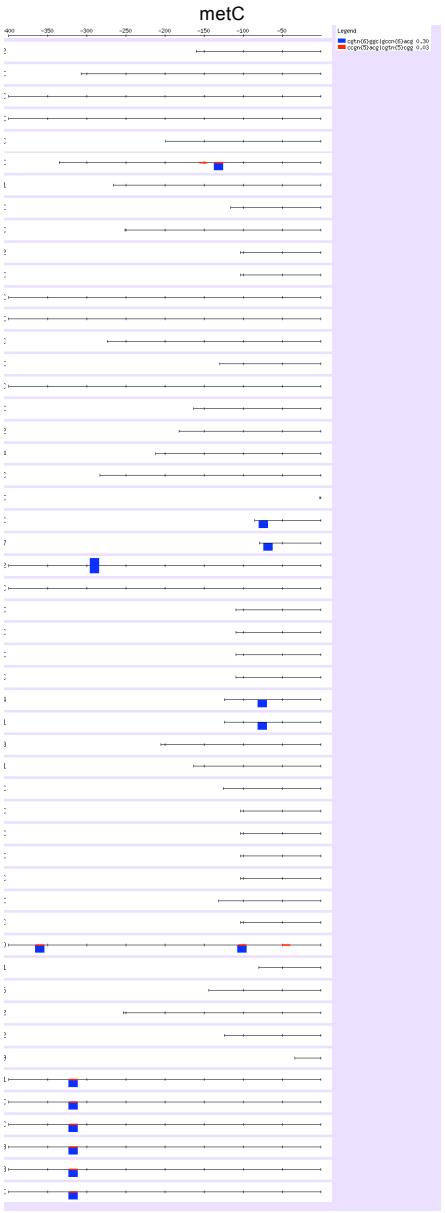


Phylogenetic footprints
MET genes in Gammaproteobacteria



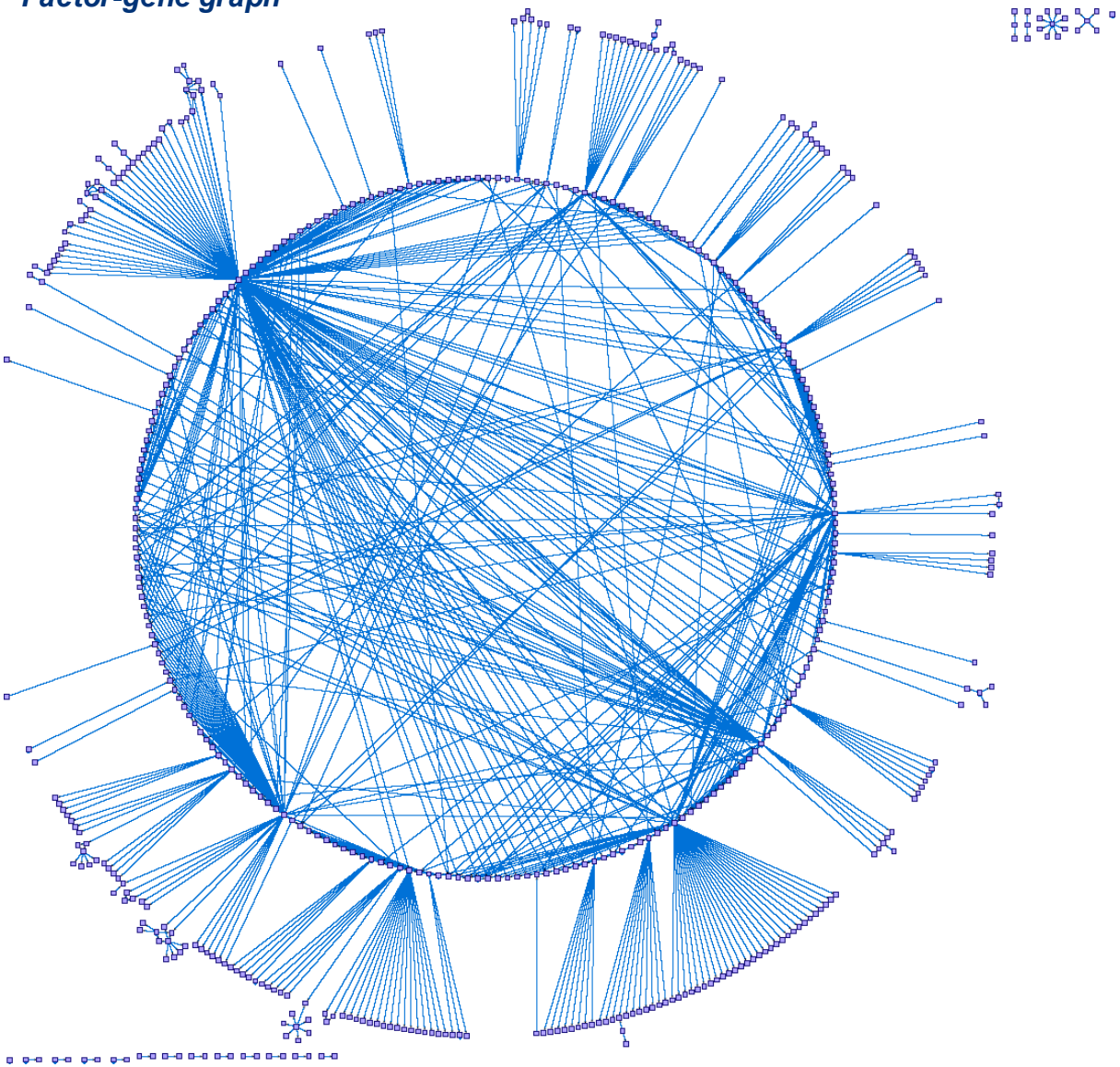






RegulonDB factor -> gene network

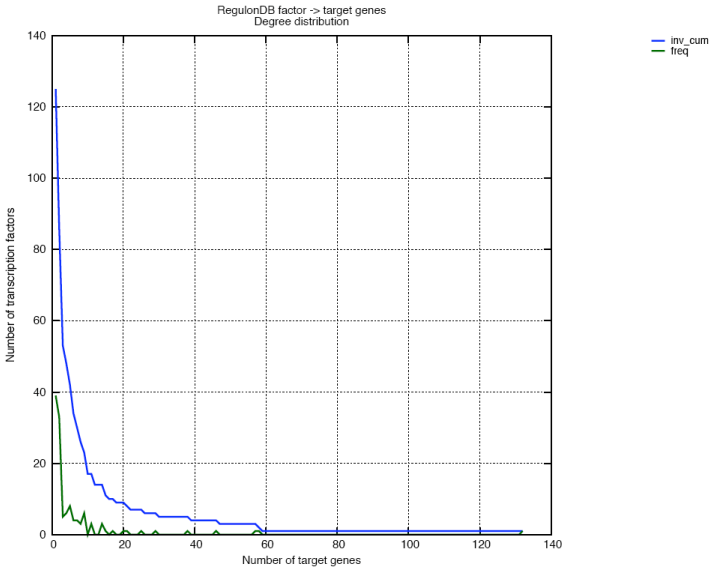
Factor-gene graph

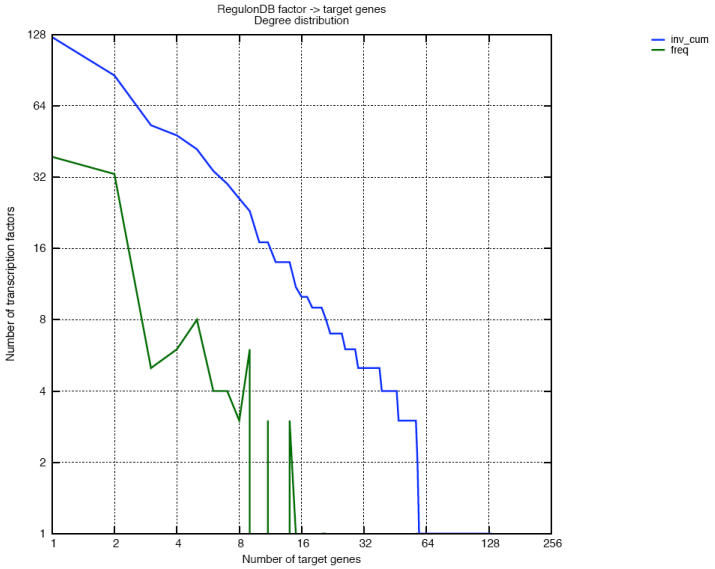


- RegulonDB (Oct. 2005 version)
- The graph represents the relationships between factors and their target genes (factor -> gene graph)
  - 125 transcription factors
  - 467 target genes
  - 847 factor->gene interactions
  - 45 self-regulations
  - Note: CRP alone regulates 132 target genes.

Number of target genes per transcription factor

Factor	Targets
CRP	132
FIS	58
IHF	57
FNR	46
ARCA	38
H-NS	29
LRP	25
FUR	21
NARL	20
SOXS	17
GADE	15
PURR	14
NTRC	14
MARA	14
GADX	11
FRUR	11
CPXR	11
PHOB	9
OXYR	9
OMPR	9
LEXA	9
FLHD	9
ARGR	9
TYRR	8
FADR	8
CYTR	8
ROB	7
NAC	7
METJ	7
CYSB	7
NAGC	6
MODE	6
GNTR	6
DNAA	6
TRPR	5
TORR	5

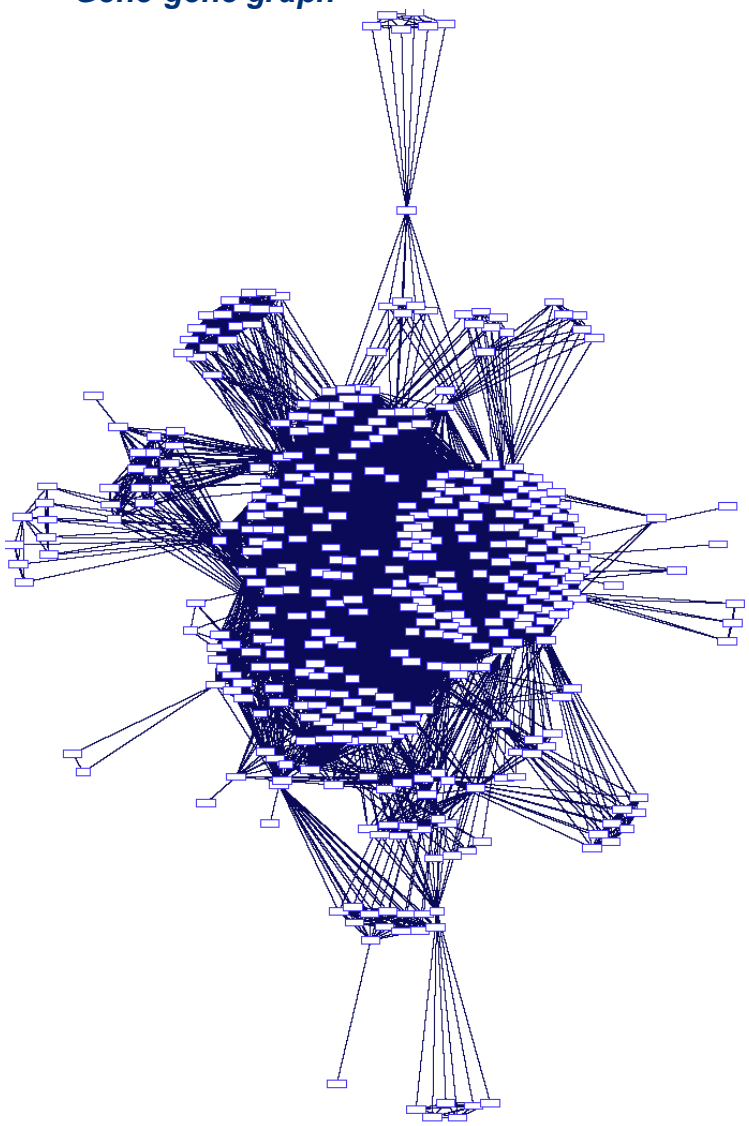


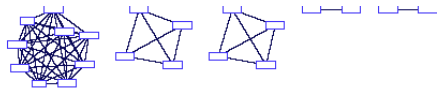


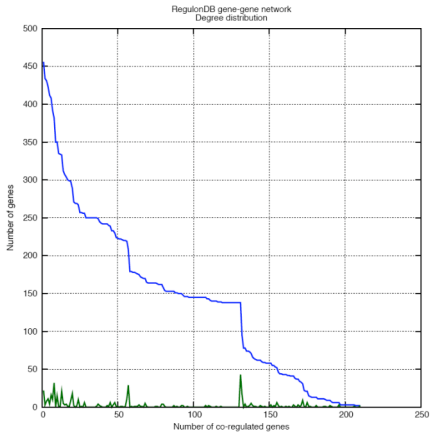
- The number of target genes per factor (out-degree) shows important variations:
  - ***Specific factors:*** most transcription factors have a very few target genes (typically 1 to 3)
  - ***General factors:*** a few factors have a very large number of target genes.
- The log-scale representation shows that the distribution of degree more or less follows a power law.

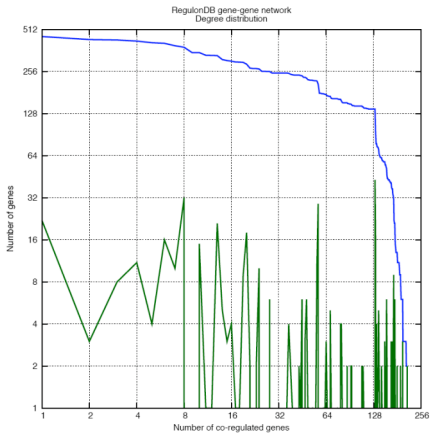
RegulonDB gene <-> gene network

Gene-gene graph





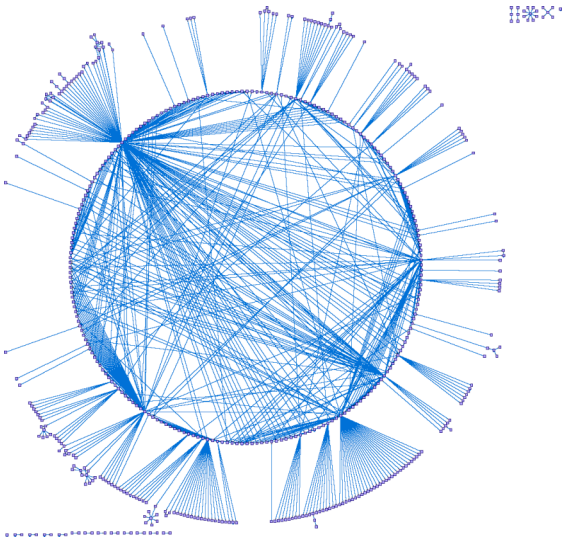




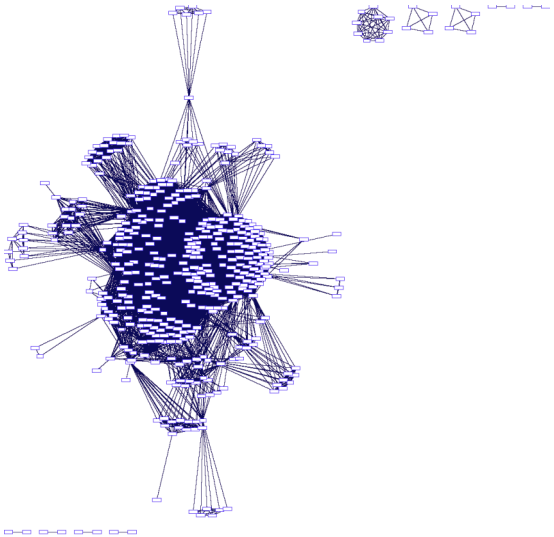
- The factor -> gene graph can be converted in a gene <-> gene graph, with an edge joining each pair of co-regulated genes (bottom).
- The gene-gene co-regulation graph contains
  - 458 nodes (genes).
    - Note: some genes are the only target of a given factor, they are thus not present in this graph.
  - 14,939 arcs (co-regulation between a pair of genes).
- Note
  - (132*131)/2=8,646 of these arcs correspond to links via CRP.
  - This network ***does not*** follow a power-law: there are many genes having many neighbours (all those regulated by some general transcription factor)

RegulonDB networks

Factor-gene graph



Gene-gene graph



- RegulonDB (Oct. 2005 version)
- The top figure displays the interaction graph between factors and genes
  (factor -> gene graph)
  - 125 transcription factors
  - 467 target genes
  - 847 factor->gene interactions
  - 45 self-regulations
  - Note: CRP alone regulates 132 target genes.
- The factor -> gene graph can be converted in a gene <-> gene graph, with an edge joining each pair of co-regulated genes (bottom).
- The gene-gene co-regulation graph contains
  - 458 nodes (genes).
    - Note: some genes are the only target of a given factor, they are thus not present in this graph.
  - 14,939 arcs (co-regulation between a pair of genes).
  - Note : (132*131)/2=8,646 of these arcs correspond to links via CRP.

Testing the principle with RegulonDB

Gene	Factor	GI
metN	MetJ	1786398
ahpC	MetJ	1786822
metK	MetJ	1789311
metC	MetJ	1789383
metB	MetJ	1790375
metF	MetJ	1790377
metA	MetJ	1790443
glyA	MetR	1788902
metH	MetR	1790450
mhpA	MhpR	1786543
purE	PurR	1786734
pyrD	PurR	1787177
pyrC	PurR	1787301
hflD	PurR	1787377
purR	PurR	1787948
cvpA	PurR	1788652
purC	PurR	1788820
purM	PurR	1788845
guaB	PurR	1788855
glyA	PurR	1788902
glnB	PurR	1788904
gcvT	PurR	1789272
purH	PurR	1790439
purL	PurR	48994899
putA	PutA	1787250
…	…	…

- Test
  - Discover motifs in upstream sequences of its orthologous genes of each gene annotated in regulonDB.
  - Link genes with similar discovered motifs.
  - Compare the inferred links with the regulon membership
- Small test case
  - Genes regulated by methionine (MetJ and MetR targets), purine (PurR targets) and proline (PutR, PutA targets)
- Complete test case
  - All the genes annotated as target gene or TF-coding in regulonDB (537 genes)

Dyad-profiles

- We can collect the significant dyads found for each gene, and display this information as “dyad profiles”:
  - Each row represents one gene
  - Each column represents one dyad found significant in at least one gene
  - Dots indicate that a dyad was not significant for a given gene (occ_sig < 0)
  - Note that the majority of the cells are empty.

	aaan{0}aat	aagn{0}ggg	aagn{1}gga	aagn{1}tgc	aagn{2}gag	aagn{2}gca	aagn{3}agc	aagn{4}gcc	aagn{6}cca	acgn{0}caa	acgn{0}tct	acgn{1}cta	acgn{1}ttg	acgn{2}aac	acgn{2}tgc	acgn{3}acg	acgn{4}cgt	actn{9}gag	agan{10}tct	agan{1}gtc	agan{5}aga	…
META	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.7	2.7	…
METB	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2.6	2.2	.	…
METF	.	.	.	.	.	0.0	.	.	.	.	.	0.5	.	.	.	.	.	.	.	.	.	…
METH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	…
METJ	.	.	.	.	.	.	.	.	.	.	2.7	2.1	.	.	.	.	.	.	1.4	3.5	.	…
METK	.	.	.	.	.	.	.	.	.	.	.	5.5	.	.	.	.	.	.	.	.	.	…
METN	.	.	.	.	.	.	.	.	.	.	.	0.5	.	.	.	.	.	.	.	.	.	…
METR	.	.	.	.	.	.	.	.	.	.	.	0.1	.	.	.	.	.	.	.	.	.	…
PURA	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	…
PURC	.	.	1.2	.	0.9	.	0.9	0.7	0.1	.	.	.	.	.	.	.	.	0.6	.	.	.	…
PURE	.	.	.	.	.	.	.	.	.	.	.	.	0.1	.	1.5	.	0.1	.	.	.	.	…
PURH	.	.	.	.	.	.	.	.	.	.	.	.	.	0.2	.	0.9	2.5	.	.	.	.	…
PURL	.	.	.	.	.	.	.	.	.	1.0	.	.	.	.	.	.	.	.	.	.	.	…
PURM	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	…
PURR	.	0.2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	…
PUTA	.	.	.	0.6	.	4.2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	…
PUTP	1.8	.	.	.	.	2.1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Compare-profiles

- The program compare-profiles was originally designed to compare binary profiles.
- It takes as input a profile file and compares each pair of profiles Q and R.
  - Profile intersection (QR),
  - union (QvR),
  - differences (Q!R, R!Q),
  - common exclusion (!Q!R)
  - Jaccard similarity = intersection/union
  - Significance of the intersection QR
    - hypergeometric significance test on the right tail : P-value + E-value + sig
  - Significance of the common exclusion !Q!R
    - hypergeometric significance test on the left tail : P-value + E-value + sig
  - Mutual information
- This program can be used to detect co-occurrence or mutual exclusion of genes across phylogenetic profiles.
- We can also use it to compare dyad profiles
  - Convert significance scores to a Boolean value

Converting dyad profiles to dyad classes

- Analysis of RegulonDB (537 genes)
  - 422 genes with at least one motif
  - 4944 dyads found significant in at least one gene
  - The profile table contains 422 rows x 4944 columns = 2,086,368 cells.
  - No more than 6,709 of these 2,086,368 cells are non-empty.
- Complete analysis of Escherichia coli (4,200 genes)
  - 2,844 genes with at least one significant motif
  - 2,844 x 17,153 dyads = 48,783,132 cells
  - No more than 33,370 of these 48,783,132 cells are non-empty.
- The profile profile matrices are very sparse.
- A (much) cheapest structure to store the same information is a 3 column file
  - dyad
  - gene
  - significance score

dyad	gene	occ_sig
gacn{0}gtc	META	0.16
agan{1}gtc	META	0.65
atcn{5}gtc	META	0.75
agan{5}aga	META	2.74
agan{1}gtc	METB	2.17
agan{10}tct	METB	2.56
gacn{0}gtc	METB	2.96
aagn{2}gca	METF	0.01
caan{3}gca	METF	0.35
gcan{0}agc	METF	0.35
acgn{1}cta	METF	0.52
agcn{2}caa	METF	0.64
gacn{0}gtc	METF	1.64
tagn{5}taa	METF	1.95
agcn{3}aag	METF	2.15
tgan{7}tca	METF	2.83
agcn{1}gca	METF	3.02
tgan{7}tca	METH	0.11
agan{6}gac	METJ	0.15
agan{9}gtc	METJ	0.15
cgtn{0}cta	METJ	0.48
cgtn{2}aga	METJ	1.04
agan{10}tct	METJ	1.35
gacn{2}cta	METJ	1.9
acgn{1}cta	METJ	2.05
acgn{0}tct	METJ	2.73
gacn{0}gtc	METJ	3.1
agan{1}gtc	METJ	3.46
taan{0}aaa	METK	0.1
gacn{0}gtc	METK	0.44
cgtn{2}aga	METK	0.51
...	...	...

Compare-classes

- The program compare-classes was designed to compare clusters
  - Clusters of co-expressed genes versus GO classes
  - Clusters of co-expressed genes versus annotated regulons
  - Annotated regulons versus annotated regulons (to detect synergy between transcription factors)
  - Chip-on-chip data versus chip-on-chip data

- We will now use compare-classes to compare the sets of dyads between genes.
- For each pair of genes Q and R, compare-classes calculates
  - number of dyads found significant in Q
  - number of dyads found significant in R
  - number of dyads in the intersection (QR)
  - number of dyads in the union (QvR)
  - Jaccard similarity = intersection/union
  - dot product (see next slide)
  - significance test on this intersection (hypergeometric, right tail)
    - P-value, E-value, sig

Gene pairs with all the RegulonDB genes
Top scoring by dot product

Rank	gene R	gene Q	R	Q	QR	QvR	jac_sim	dotprod	P_val	E_val	sig
1	PUTP	PUTA	38	37	25	50	0.5	2542.1	7.10E-49	6.30E-44	43.2
2	RECA	LEXA	57	101	14	144	0.09722	1889.8	3.30E-12	2.90E-07	6.53
3	GLNL	GLNA	51	88	32	107	0.29907	1713.11	7.00E-46	6.20E-41	40.21
4	UDP	MTLA	15	54	11	58	0.18966	1043.88	1.20E-19	1.10E-14	13.98
5	UVRD	LEXA	41	101	11	131	0.08397	831.49	2.80E-10	2.50E-05	4.6
6	TYRP	AROF	25	32	11	46	0.23913	815.729	5.10E-19	4.50E-14	13.34
7	PUTP	GLNA	38	88	3	123	0.02439	628.522	0.02936	2608.27	-3.42
8	MTLA	MLC	54	11	5	60	0.08333	448.632	5.70E-08	0.00503	2.3
9	PSPF	PSPA	64	64	64	64	1	437.011	######	6.40E-143	142.19
10	PUTA	GLNA	37	88	3	122	0.02459	431.504	0.02739	2432.93	-3.39
11	ILVY	ILVC	93	75	51	117	0.4359	386.541	9.00E-77	8.00E-72	71.1
12	SULA	RECA	41	57	11	87	0.12644	358.099	4.20E-13	3.70E-08	7.43
13	UVRD	RECA	41	57	10	88	0.11364	347.532	1.60E-11	1.40E-06	5.86
14	UVRA	LEXA	12	101	5	108	0.0463	346.642	2.30E-06	0.20236	0.69
15	PUTP	PDHR	38	59	1	96	0.01042	330.398	0.36741	32637.392	-4.51
16	POLA	DNAA	39	32	15	56	0.26786	296.35	6.80E-25	6.10E-20	19.22
17	BIOB	BIOA	31	36	22	45	0.48889	287.556	4.70E-44	4.20E-39	38.38
18	METJ	METB	45	23	19	49	0.38776	286.685	1.70E-37	1.50E-32	31.81
19	MTLA	CYDA	54	19	3	70	0.04286	281.577	0.00105	93.699	-1.97
20	PUTA	PDHR	37	59	1	95	0.01053	269.359	0.35971	31953.518	-4.5
21	NAGE	MTLA	19	54	6	67	0.08955	227.65	3.10E-08	0.00276	2.56
22	UDP	MLC	15	11	5	21	0.2381	200.044	5.60E-11	5.00E-06	5.3
23	MTLA	ADHE	54	8	3	59	0.05085	197.458	6.60E-05	5.896	-0.77
24	MALK	MALE	21	73	19	75	0.25333	197.254	2.70E-34	2.40E-29	28.62
25	NRDD	NRDA	39	63	20	82	0.2439	192.165	2.60E-29	2.30E-24	23.63
26	YFID	MTLA	29	54	5	78	0.0641	183.569	1.30E-05	1.113	-0.05
27	YFID	CYDA	29	19	6	42	0.14286	165.44	6.10E-10	5.40E-05	4.27
28	SULA	LEXA	41	101	4	138	0.02899	164.636	0.00931	826.926	-2.92
29	MTLA	CDD	54	14	4	64	0.0625	162.407	1.20E-05	1.043	-0.02
30	UVRA	RECA	12	57	5	64	0.07812	160.535	1.30E-07	0.01127	1.95
31	SSB	LEXA	22	101	2	121	0.01653	159.836	0.07326	6507.387	-3.81
32	CUER	COPA	43	47	29	61	0.47541	157.892	2.40E-53	2.20E-48	47.67
33	UHPT	MTLA	12	54	3	63	0.04762	157.165	0.00025	22.457	-1.35
34	MTLA	AGAR	54	31	8	77	0.1039	154.984	7.70E-10	6.80E-05	4.17
35	PUTP	GLNL	38	51	2	87	0.02299	150.447	0.05794	5147.255	-3.71
36	TYRR	AROF	29	32	6	55	0.10909	147.935	1.90E-08	0.0017	2.77
37	XSEA	GUAB	29	37	13	53	0.24528	139.955	1.30E-21	1.20E-16	15.92
38	ULAA	MTLA	27	54	3	78	0.03846	132.93	0.00299	265.909	-2.42
39	MTLA	GLPA	54	15	5	64	0.07812	130.283	3.60E-07	0.03161	1.5
40	PSTS	PHOB	28	31	7	52	0.13462	130.01	2.00E-10	1.80E-05	4.75
41	GLNK	GLNA	14	88	6	96	0.0625	125.801	7.20E-08	0.00637	2.2
42	MALP	MALE	19	73	11	81	0.1358	124.547	2.30E-16	2.00E-11	10.69
43	ARGE	ARGC	38	24	14	48	0.29167	122.146	3.10E-25	2.70E-20	19.56
44	PDHR	GLCC	59	13	7	65	0.10769	118.735	3.90E-11	3.40E-06	5.46
45	TYRR	TYRP	29	25	6	48	0.125	114.252	3.90E-09	0.00034	3.47
46	METN	FPR	79	24	13	90	0.14444	113.54	3.40E-18	3.10E-13	12.51
47	MTLA	MALE	54	73	8	119	0.06723	112.418	9.20E-07	0.08208	1.09
48	PDHR	FADB	59	14	5	68	0.07353	110.459	3.80E-07	0.03338	1.48
49	TREB	MTLA	36	54	3	87	0.03448	107.764	0.00682	605.69	-2.78
50	PUTA	GLNL	37	51	2	86	0.02326	103.141	0.05525	4907.985	-3.69

Dot product

$$dp = \sum_{i=1}^{p} \left(x_{Ai} \cdot x_{Bi} \right)$$

Jaccard similarity

$$S_J = \frac{(Q \cap R)}{(Q \cup R)}$$

Hypergeometric significance

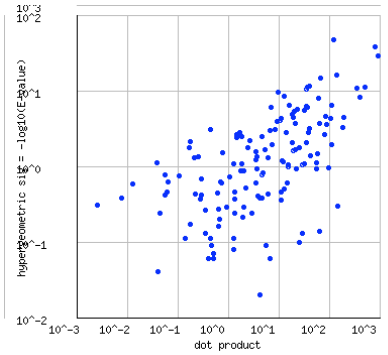
$$P(X = x) = \frac{C_m^x C_n^{k-x}}{C_{m+n}^k}$$

$$Pval = P(X \geq x) = \sum_{i=x}^{\min(k,m)} \frac{C_m^i C_n^{k-i}}{C_{m+n}^k}$$

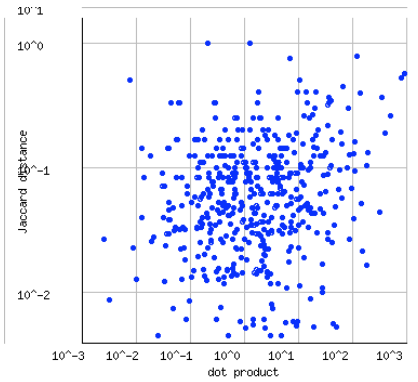
$$Eval = Pval \cdot N_{test}$$

$$sig = -\log_{10}(Eval)$$

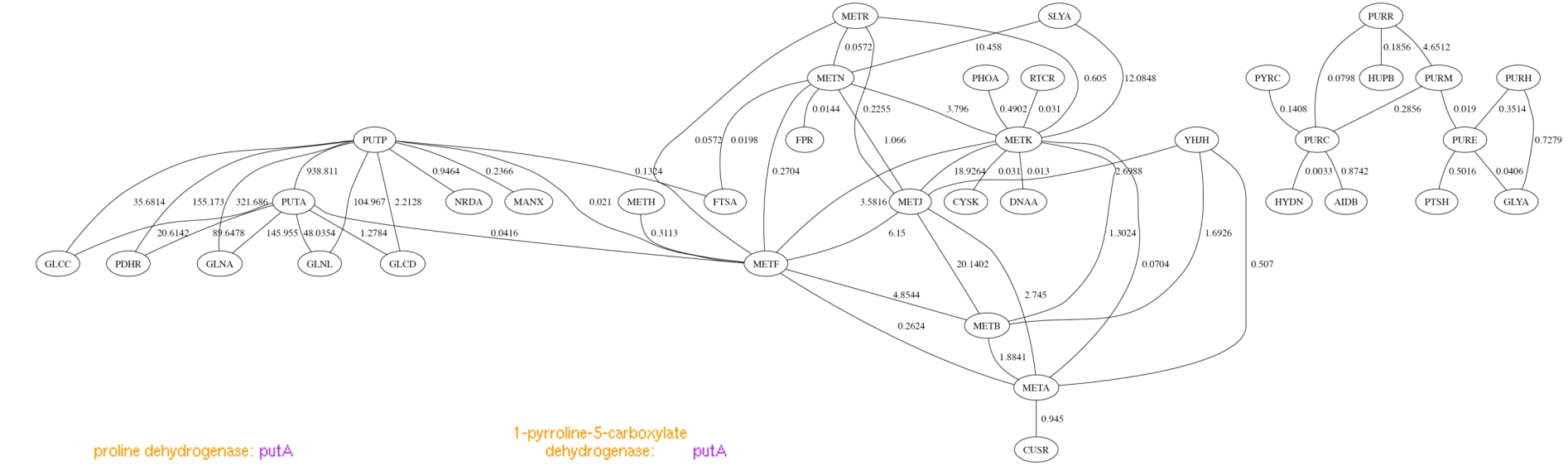
Dot product versus hypergeometric (log axes)

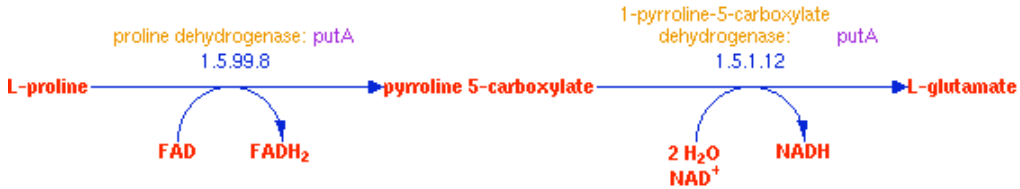


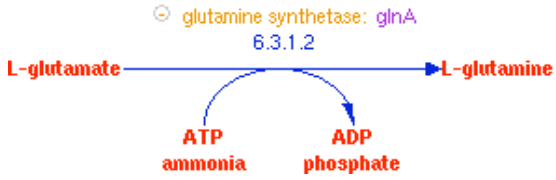
Dot product versus Jaccard (log axes)



Inferred co-regulation graph (selection)

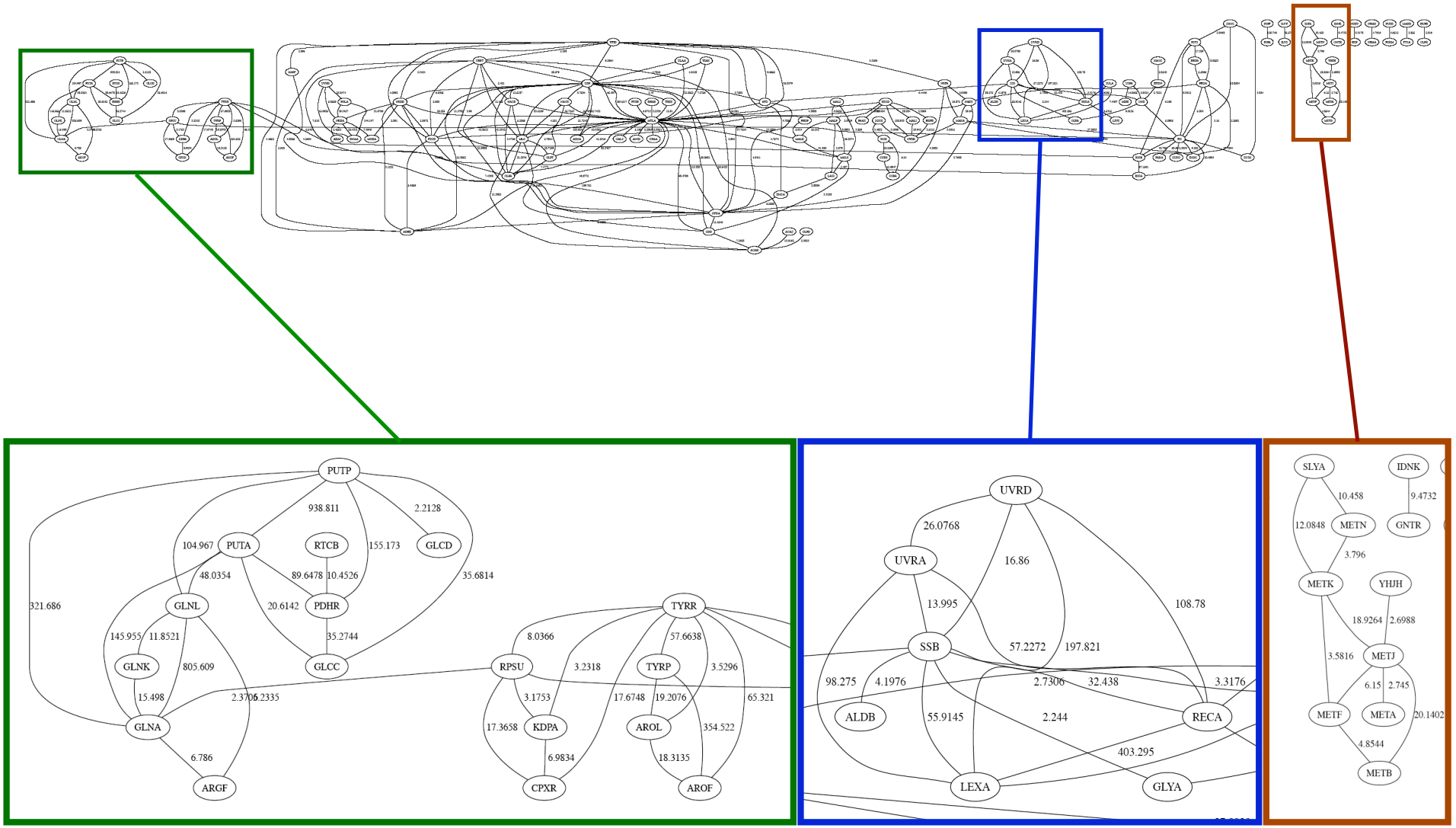




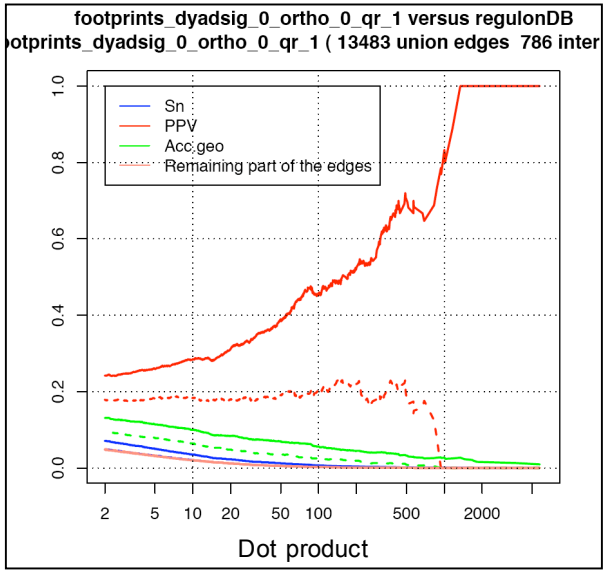


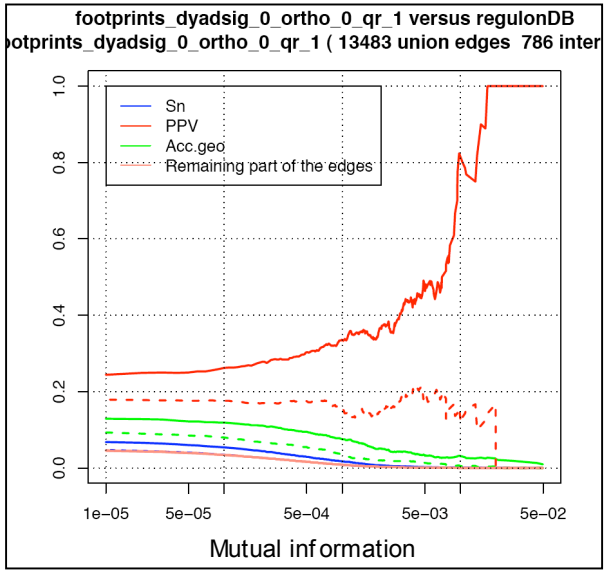
Metabolic pathways from http://biocyc.org/

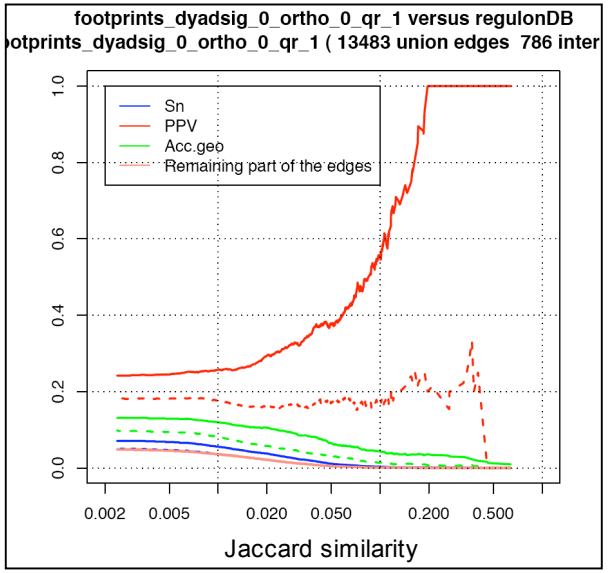
All the pairs of dyad profiles with dp >= 2

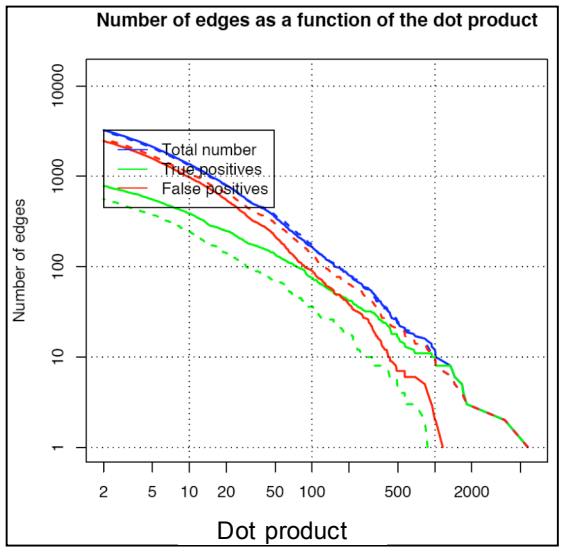


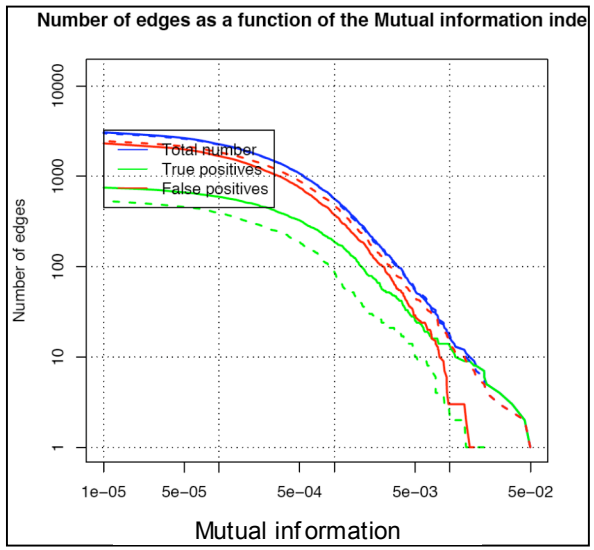
Validation of inferred co-regulation network

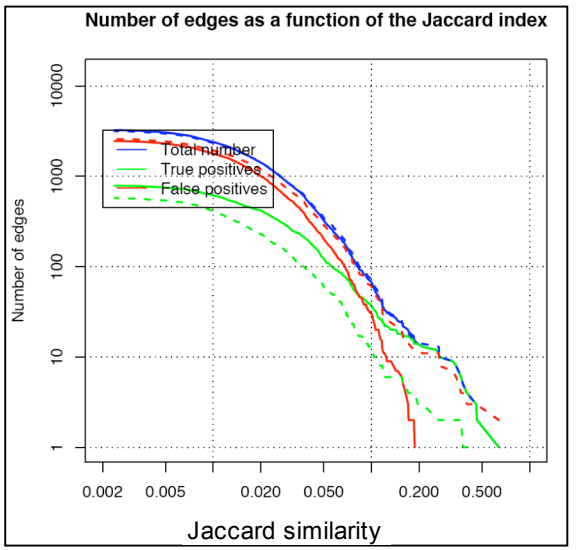












Predicted versus annotated gene pairs

Gene pair	Annot	Score	Sum pred	Sum annot	P.and.A	P.not.A	A.not.P	PPV	Sn
PUTP_PUTA	0	2542.1	1	0	0	1	17076	0.000	0.000
RECA_LEXA	1	1889.8	2	1	1	1	17075	0.500	0.000
GLNL_GLNA	1	1713.1	3	2	2	1	17074	0.667	0.000
UDP_MTLA	1	1043.9	4	3	3	1	17073	0.750	0.000
UVRD_LEXA	1	831.5	5	4	4	1	17072	0.800	0.000
TYRP_AROF	1	815.7	6	5	5	1	17071	0.833	0.000
PUTP_GLNA	1	628.5	7	6	6	1	17070	0.857	0.000
MTLA_MLC	1	448.6	8	7	7	1	17069	0.875	0.000
PSPF_PSPA	0	437.0	9	7	7	2	17069	0.778	0.000
PUTA_GLNA	0	431.5	10	7	7	3	17069	0.700	0.000
ILVY_ILVC	1	386.5	11	8	8	3	17068	0.727	0.001
SULA_RECA	1	358.1	12	9	9	3	17067	0.750	0.001
UVRD_RECA	1	347.5	13	10	10	3	17066	0.769	0.001
UVRA_LEXA	1	346.6	14	11	11	3	17065	0.786	0.001
PUTP_PDHR	1	330.4	15	12	12	3	17064	0.800	0.001
POLA_DNAA	1	296.4	16	13	13	3	17063	0.813	0.001
BIOB_BIOA	1	287.6	17	14	14	3	17062	0.824	0.001
METJ_METB	0	286.7	18	14	14	4	17062	0.778	0.001
MTLA_CYDA	0	281.6	19	14	14	5	17062	0.737	0.001
PUTA_PDHR	0	269.4	20	14	14	6	17062	0.700	0.001
NAGE_MTLA	1	227.7	21	15	15	6	17061	0.714	0.001
UDP_MLC	1	200.0	22	16	16	6	17060	0.727	0.001
MTLA_ADHE	1	197.5	23	17	17	6	17059	0.739	0.001
MALK_MALE	1	197.3	24	18	18	6	17058	0.750	0.001
NRDD_NRDA	0	192.2	25	18	18	7	17058	0.720	0.001
YFID_MTLA	0	183.6	26	18	18	8	17058	0.692	0.001
YFID_CYDA	1	165.4	27	19	19	8	17057	0.704	0.001
SULA_LEXA	1	164.6	28	20	20	8	17056	0.714	0.001
MTLA_CDD	1	162.4	29	21	21	8	17055	0.724	0.001
UVRA_RECA	1	160.5	30	22	22	8	17054	0.733	0.001
SSB_LEXA	1	159.8	31	23	23	8	17053	0.742	0.002
CUER_COPA	0	157.9	32	23	23	9	17053	0.719	0.002
UHPT_MTLA	1	157.2	33	24	24	9	17052	0.727	0.002
MTLA_AGAR	0	155.0	34	24	24	10	17052	0.706	0.002
PUTP_GLNL	0	150.4	35	24	24	11	17052	0.686	0.002
TYRR_AROF	1	147.9	36	25	25	11	17051	0.694	0.002
XSEA_GUAB	1	140.0	37	26	26	11	17050	0.703	0.002
ULAA_MTLA	1	132.9	38	27	27	11	17049	0.711	0.002
MTLA_GLPA	1	130.3	39	28	28	11	17048	0.718	0.002
PSTS_PHOB	1	130.0	40	29	29	11	17047	0.725	0.002
GLNK_GLNA	1	125.8	41	30	30	11	17046	0.732	0.002
MALP_MALE	1	124.5	42	31	31	11	17045	0.738	0.002
ARGE_ARGC	1	122.1	43	32	32	11	17044	0.744	0.002
PDHR_GLCC	0	118.7	44	32	32	12	17044	0.727	0.002
TYRR_TYRP	1	114.3	45	33	33	12	17043	0.733	0.002
METN_FPR	0	113.5	46	33	33	13	17043	0.717	0.002
MTLA_MALE	1	112.4	47	34	34	13	17042	0.723	0.002
PDHR_FADB	0	110.5	48	34	34	14	17042	0.708	0.002
TREB_MTLA	1	107.8	49	35	35	14	17041	0.714	0.002

- Results with 536 genes annotated in RegulonDB
  - Among the 537*536/2=~143,916 gene pairs considered here (between 537 genes annotated in regulonDB), 2,790 show a common motif.
  - Among the 50 top-scoring predicted gene pairs, 35 are annotated as co-regulated in RegulonDB.
  - Among the 15 top-scoring predicted which are non-annotated, at least 8 are actually involved in related pathways
    - Some are already known to be co-regulated, but the annotation is not yet entered in RegulonDB.
    - Some others are very likely to be co-regulated.
- Whole genome analysis (4200 genes)
  - Among the ~16M possible gene pairs 54,099 show a common motif

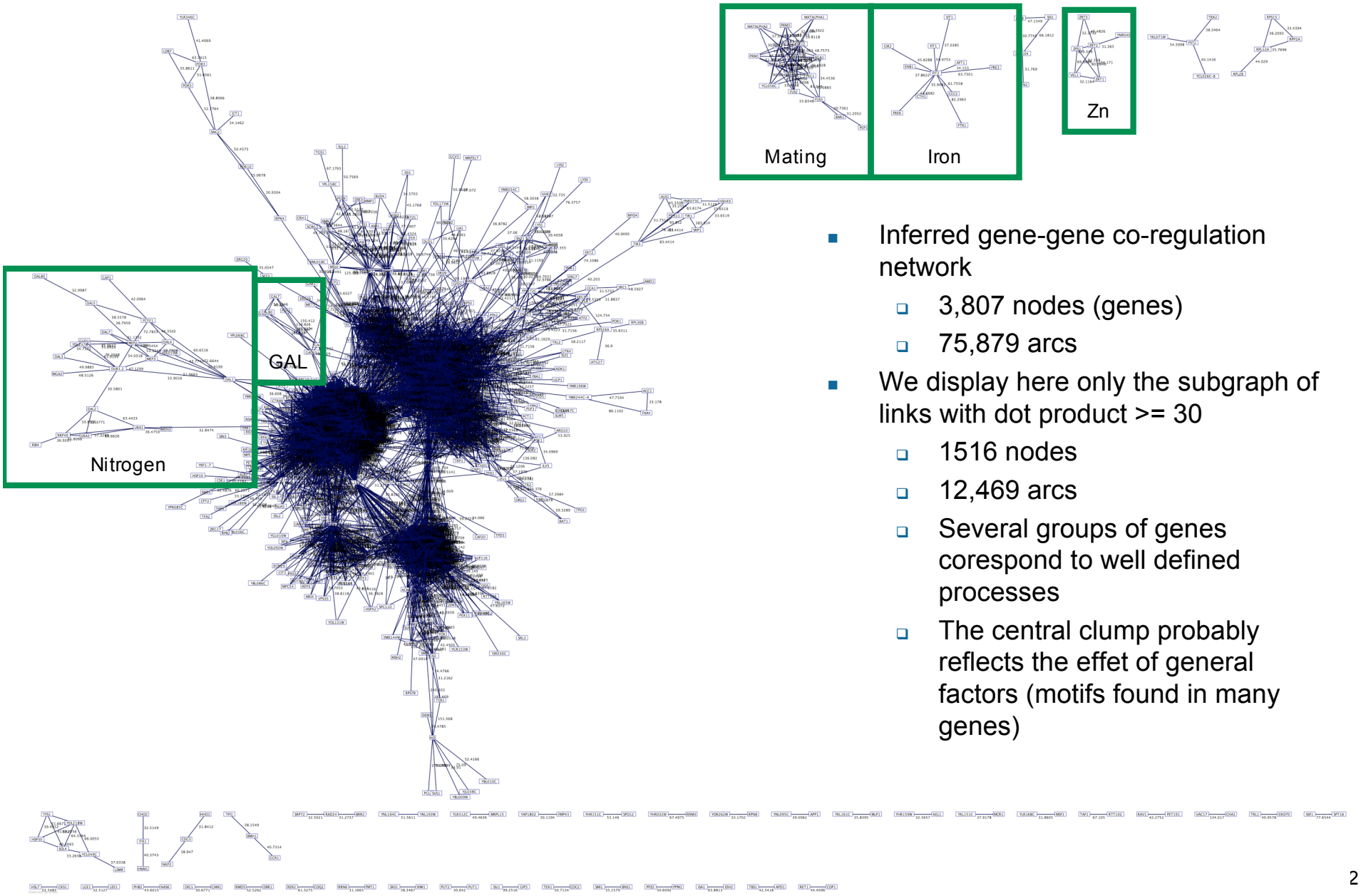
Regulon-wise sensitivity for Escherichia coli K12

Regulon	Nodes	Max edges	Predicted edges	Extra edges	Sn	Rand ER Sn	Rand ER extra edges	Rand ND Sn	Rand ND extra edges
CRP	405	81810	3400	8970	0.04	0.04	9805.00	0.03	8956.50
FNR	265	34980	1412	7098	0.04	0.03	7566.00	0.02	6682.00
IHF	207	21321	802	6019	0.04	0.03	6243.00	0.03	5635.50
Fis	166	13695	209	3438	0.02	0.01	3522.00	0.01	2862.00
ArcA	151	11325	842	5614	0.07	0.06	6037.00	0.04	4925.50
ArcA	151	11325	842	5614	0.07	0.06	6037.00	0.04	4925.50
H-NS	113	6328	104	2425	0.02	0.01	2502.00	0.01	2091.50
NarL	98	4753	220	2578	0.05	0.02	2815.00	0.03	3087.00
Lrp	88	3828	87	1712	0.02	0.01	1802.00	0.01	1585.50
FlhDC	84	3486	134	1784	0.04	0.02	1937.00	0.03	2705.00
Fur	77	2926	287	2232	0.10	0.03	2620.00	0.03	2585.00
CpxR	55	1485	80	1359	0.05	0.02	1471.00	0.03	1861.00
ModE	46	1035	77	639	0.07	0.01	781.00	0.03	1542.50
NtrC	44	946	95	358	0.10	0.00	543.00	0.01	1228.50
NarP	40	780	42	907	0.05	0.01	973.00	0.02	1120.00
PhoB	35	595	40	251	0.07	0.00	330.00	0.02	1080.50
FruR	33	528	119	2788	0.23	0.17	2846.00	0.03	1135.50
PhoP	31	465	12	499	0.03	0.01	516.00	0.01	975.50
PurR	31	465	73	1293	0.16	0.05	1394.00	0.04	1045.00
FhlA	30	435	93	225	0.21	0.01	406.00	0.03	926.00
ArgR	28	378	123	2329	0.33	0.16	2456.00	0.04	963.50
LexA	28	378	95	883	0.25	0.03	1047.00	0.02	885.00
GadE	27	351	28	1126	0.08	0.05	1147.00	0.03	889.00
IscR	26	325	30	727	0.09	0.02	774.00	0.02	809.50
SoxS	25	300	12	743	0.04	0.03	751.00	0.01	617.00
CysB	24	276	13	256	0.05	0.01	278.00	0.02	766.50
RcsAB	22	231	9	458	0.04	0.01	473.00	0.01	680.00
MarA	21	210	9	831	0.04	0.04	834.00	0.02	486.00
NagC	18	153	26	928	0.17	0.06	962.00	0.00	592.00
OxyR	18	153	11	470	0.07	0.01	488.00	0.01	463.50
GadX	17	136	7	780	0.05	0.03	786.00	0.01	416.50
Nac	15	105	9	826	0.09	0.06	831.00	0.01	449.00
MetJ	13	78	31	494	0.40	0.05	548.00	0.04	507.50
OmpR	13	78	15	762	0.19	0.08	779.00	0.03	433.50
Rob	13	78	5	468	0.06	0.03	474.00	0.01	367.50
CytR	12	66	3	346	0.05	0.01	351.00	0.08	487.00
GntR	12	66	21	356	0.32	0.01	397.00	0.02	327.00
HyfR	12	66	5	36	0.08	0.00	46.00	0.01	296.00
PaaX	12	66	10	74	0.15	0.00	94.00	0.03	435.50
TrpR	12	66	12	567	0.18	0.04	586.00	0.01	471.50
AgaR	11	55	23	738	0.42	0.09	774.00	0.02	343.00
FadR	11	55	11	230	0.20	0.00	252.00	0.04	350.00
TorR	11	55	4	91	0.07	0.00	99.00	0.02	357.00
TyrR	11	55	26	354	0.47	0.04	402.00	0.00	423.50
AppY	10	45	5	40	0.11	0.00	50.00	0.01	329.00
CaiF	10	45	16	304	0.36	0.00	336.00	0.06	399.50
DnaA	10	45	16	987	0.36	0.28	994.00	0.03	337.50
MalT	10	45	17	293	0.38	0.03	324.00	0.03	440.00
AllR	9	36	0	47	0.00	0.00	47.00	0.00	117.50
AraC	9	36	8	77	0.22	0.00	93.00	0.03	333.50
BaeR	9	36	9	32	0.25	0.00	50.00	0.01	308.50
Cbl	9	36	4	20	0.11	0.00	28.00	0.01	259.50
CdaR	9	36	6	18	0.17	0.00	30.00	0.03	344.00
CsgD	9	36	8	323	0.22	0.03	337.00	0.01	356.00
DgsA	9	36	17	821	0.47	0.24	838.00	0.01	328.50
GalR	9	36	8	131	0.22	0.01	146.00	0.00	324.50
GalS	9	36	8	131	0.22	0.01	146.00	0.00	324.50
GlpR	9	36	16	849	0.44	0.17	869.00	0.03	335.00
YiaJ	9	36	1	157	0.03	0.00	159.00	0.00	86.00
ExuR	8	28	7	61	0.25	0.00	75.00	0.02	201.50
HU	8	28	4	122	0.14	0.00	130.00	0.04	282.50
PdhR	8	28	10	956	0.36	0.25	962.00	0.04	327.00
CueR	7	21	12	51	0.57	0.00	75.00	0.05	284.50
DcuR	7	21	8	561	0.38	0.10	573.00	0.05	243.00
EnvY	7	21	11	416	0.52	0.07	435.00	0.07	293.00
FucR	7	21	6	143	0.29	0.00	155.00	0.02	177.00
GlcC	7	21	10	196	0.48	0.00	216.00	0.05	303.00
GutM	7	21	9	378	0.43	0.00	396.00	0.02	313.00
GutR	7	21	9	378	0.43	0.00	396.00	0.02	313.00
IdnR	7	21	10	201	0.48	0.00	221.00	0.02	206.00
NanR	7	21	1	58	0.05	0.00	60.00	0.00	166.50

Regulon	Nodes	Max edges	Predicted edges	Extra edges	Sn	Rand ER Sn	Rand ER extra edges	Rand ND Sn	Rand ND extra edges
PspF	7	21	14	628	0.67	0.14	650.00	0.07	307.00
TdcA	7	21	1	264	0.05	0.02	265.00	0.00	123.50
TdcR	7	21	1	264	0.05	0.02	265.00	0.00	123.50
UlaR	7	21	21	676	1.00	0.10	714.00	0.07	281.00
UxuR	7	21	3	25	0.14	0.00	31.00	0.00	128.50
AlsR	6	15	0	22	0.00	0.00	22.00	0.00	52.00
ChbR	6	15	3	45	0.20	0.00	51.00	0.00	213.00
CusR	6	15	1	46	0.07	0.00	48.00	0.00	88.50
DeoR	6	15	2	98	0.13	0.03	101.00	0.17	234.50
EvgA	6	15	3	6	0.20	0.00	12.00	0.00	231.50
GatR	6	15	10	320	0.67	0.07	338.00	0.00	214.50
HcaR	6	15	1	13	0.07	0.00	15.00	0.07	163.00
LsrR	6	15	15	391	1.00	0.07	419.00	0.03	277.50
MhpR	6	15	0	4	0.00	0.00	4.00	0.00	91.50
NhaR	6	15	3	150	0.20	0.00	156.00	0.03	260.00
NikR	6	15	10	10	0.67	0.00	30.00	0.00	210.50
RbsR	6	15	7	244	0.47	0.00	258.00	0.07	267.50
XylR	6	15	2	158	0.13	0.03	161.00	0.00	171.00
ArgP	5	10	4	547	0.40	0.35	548.00	0.00	168.50
BirA	5	10	10	60	1.00	0.00	80.00	0.00	200.50
CsiR	5	10	1	31	0.10	0.00	33.00	0.05	128.50
DicA	5	10	0	0	0.00	0.00	0.00	0.00	0.00
LeuO	5	10	6	172	0.60	0.10	182.00	0.00	228.00
MetR	5	10	4	214	0.40	0.00	222.00	0.05	163.00
NsrR	5	10	4	141	0.40	0.05	148.00	0.05	155.50
PrpR	5	10	1	44	0.10	0.00	46.00	0.00	176.00
Ada	4	6	1	0	0.17	0.00	2.00	0.00	85.50
AsnC	4	6	3	782	0.50	0.50	782.00	0.00	132.00
AtoC	4	6	0	4	0.00	0.00	4.00	0.00	37.50
BetI	4	6	6	95	1.00	0.00	107.00	0.08	171.50
CynR	4	6	0	37	0.00	0.00	37.00	0.00	40.50
DhaR	4	6	3	34	0.50	0.00	40.00	0.00	127.50
GadW	4	6	0	39	0.00	0.00	39.00	0.00	116.00
GcvA	4	6	1	181	0.17	0.00	183.00	0.00	112.00
IclR	4	6	0	307	0.00	0.00	307.00	0.00	90.00
KdpE	4	6	3	83	0.50	0.00	89.00	0.00	119.00
LrhA	4	6	1	54	0.17	0.00	56.00	0.08	165.00
QseB	4	6	2	61	0.33	0.00	65.00	0.08	169.00
RhaS	4	6	3	96	0.50	0.00	102.00	0.00	158.50
UidR	4	6	0	5	0.00	0.00	5.00	0.00	32.00
AcrR	3	3	1	97	0.33	0.00	99.00	0.00	86.50
AllS	3	3	0	10	0.00	0.00	10.00	0.00	38.00
AscG	3	3	3	3	1.00	0.00	9.00	0.00	130.50
CadC	3	3	0	141	0.00	0.00	141.00	0.17	126.00
DsdC	3	3	3	78	1.00	0.00	84.00	0.00	136.00
LacI	3	3	1	33	0.33	0.00	35.00	0.00	78.00
LldR	3	3	3	78	1.00	0.00	84.00	0.00	118.00
MalI	3	3	1	64	0.33	0.17	65.00	0.00	134.50
MarR	3	3	3	24	1.00	0.00	30.00	0.00	123.00
MelR	3	3	1	21	0.33	0.00	23.00	0.00	93.00
MngR	3	3	1	58	0.33	0.00	60.00	0.17	126.00
MprA	3	3	1	229	0.33	0.00	231.00	0.00	122.50
MtlR	3	3	1	432	0.33	0.00	434.00	0.00	122.00
NorR	3	3	3	128	1.00	0.00	134.00	0.00	129.00
SdiA	3	3	1	99	0.33	0.00	101.00	0.00	124.50
SoxR	3	3	0	20	0.00	0.00	20.00	0.00	45.00
ZraR	3	3	3	45	1.00	0.00	51.00	0.00	110.50
Zur	3	3	3	355	1.00	0.17	360.00	0.00	123.50
BolA	2	1	0	0	0.00	0.00	0.00	0.00	0.00
CspA	2	1	1	382	1.00	1.00	382.00	0.00	93.50
EbgR	2	1	1	73	1.00	0.00	75.00	0.50	83.50
FabR	2	1	1	104	1.00	0.00	106.00	0.00	83.50
HdfR	2	1	1	46	1.00	0.00	48.00	0.00	83.50
HipB	2	1	1	50	1.00	0.00	52.00	0.00	83.50
IlvY	2	1	1	293	1.00	0.50	294.00	0.00	78.00
LysR	2	1	1	142	1.00	0.00	144.00	0.00	88.50
NadR	2	1	0	11	0.00	0.00	11.00	0.00	43.50
PepA	2	1	0	376	0.00	0.00	376.00	0.00	40.00
RhaR	2	1	0	2	0.00	0.00	2.00	0.00	38.50
RtcR	2	1	0	0	0.00	0.00	0.00	0.00	0.00
TreR	2	1	1	34	1.00	0.00	36.00	0.00	97.50
XapR	2	1	0	3	0.00	0.00	3.00	0.00	38.50

Inferring co-regulation network for the yeast Saccharomyces cerevisiae

Saccharomyces cerevisiae, Saccharomycetaceae
Inferred co-regulation network

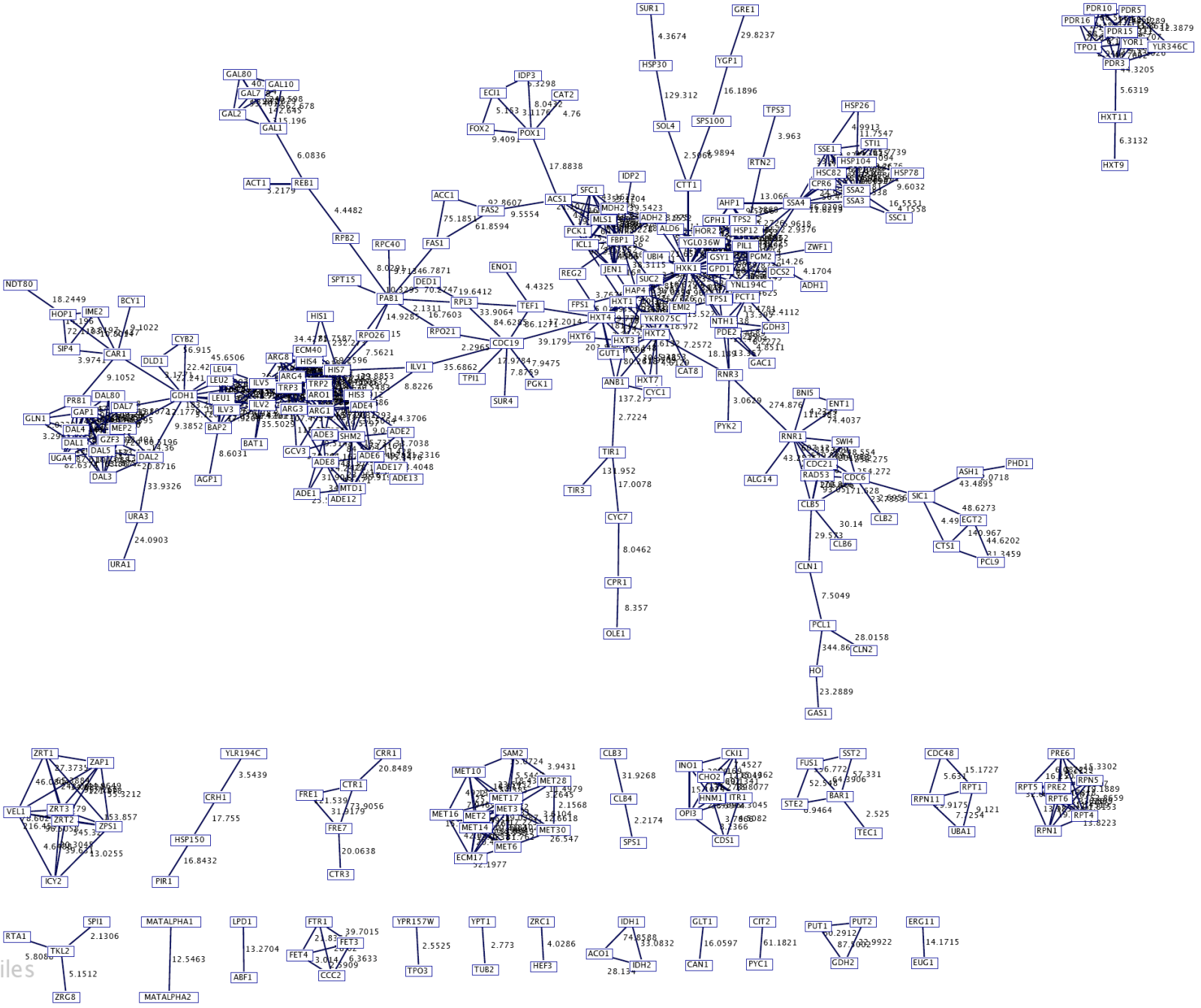


- Inferred gene-gene co-regulation network
  - 3,807 nodes (genes)
  - 75,879 arcs
- We display here only the subgraph of links with dot product >= 30
  - 1516 nodes
  - 12,469 arcs
  - Several groups of genes corespond to well defined processes
  - The central clump probably reflects the effet of general factors (motifs found in many genes)

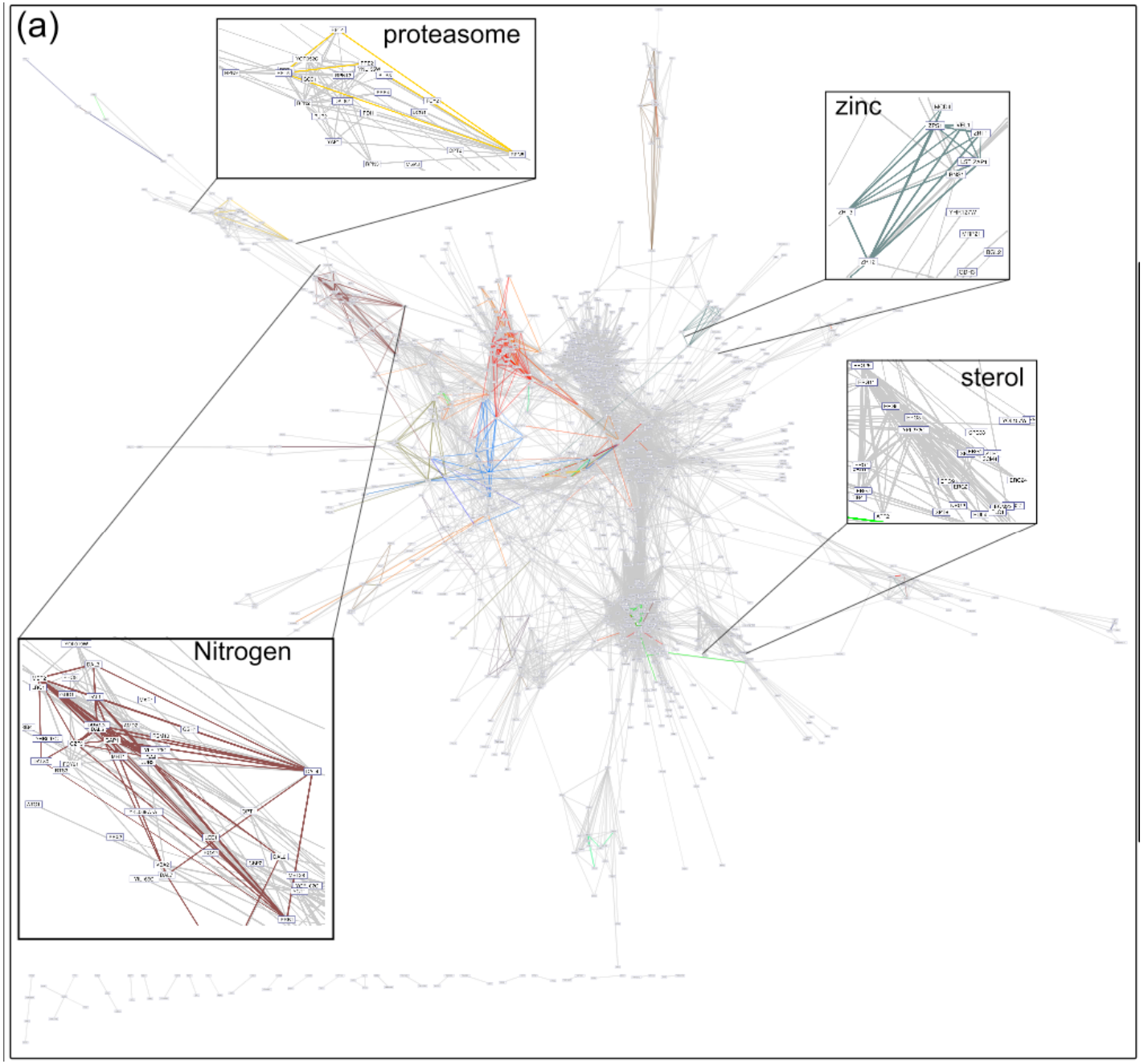
Regulon-wise sensitivity

Regulon	Nodes	Max edges	Predicted edges	Extra edges	Sn	Rand ER Sn	Rand ER extra edges	Rand ND Sn	Rand ND extra edges
MSN4	58	1653	89	1983	0.05	0.01	2128.53	0.01	1606.20
MSN2	56	1540	82	1877	0.05	0.01	2011.60	0.01	1571.57
ZAP1	54	1431	25	785	0.02	0.00	833.20	0.00	1154.13
TEC1	42	861	6	726	0.01	0.00	737.00	0.00	836.43
GCN4	40	780	101	1139	0.13	0.01	1343.60	0.01	1056.67
ABF1	37	666	12	1399	0.02	0.01	1415.60	0.01	976.53
RAP1	32	496	9	1008	0.02	0.01	1032.73	0.00	767.97
YAP1	32	496	8	423	0.02	0.00	438.07	0.00	524.50
GLN3	31	465	69	852	0.15	0.01	983.13	0.01	915.67
MIG1	26	325	37	1377	0.11	0.03	1438.47	0.01	673.77
UME6	26	325	20	519	0.06	0.00	596.40	0.01	703.60
RLM1	25	300	6	339	0.02	0.00	354.20	0.01	725.50
OAF1	24	276	10	218	0.04	0.00	239.53	0.01	637.03
HSF1	21	210	29	443	0.14	0.00	500.00	0.00	492.27
PHO2	21	210	34	506	0.16	0.00	576.13	0.01	629.03
DAL80	19	171	39	428	0.23	0.01	504.13	0.01	675.60
REB1	19	171	8	575	0.05	0.01	602.13	0.01	596.70
INO2	19	171	22	131	0.13	0.00	174.67	0.01	531.30
INO4	19	171	22	131	0.13	0.00	174.67	0.01	531.30
PIP2	19	171	10	152	0.06	0.00	173.80	0.01	458.63
GCR1	18	153	6	808	0.04	0.02	815.80	0.00	415.60
BAS1	17	136	34	425	0.25	0.00	495.73	0.01	559.40
PDR1	16	120	32	367	0.27	0.01	429.60	0.01	457.73
CBF1	16	120	23	432	0.19	0.01	477.67	0.01	499.97
MOT3	15	105	15	975	0.14	0.03	1043.13	0.02	535.20
MIG2	15	105	10	738	0.10	0.02	758.07	0.01	428.67
ROX1	15	105	5	768	0.05	0.01	819.87	0.01	462.67
HAP3	15	105	3	348	0.03	0.00	353.33	0.01	457.60
MCM1	14	91	3	480	0.03	0.01	487.67	0.01	393.80
HAP4	14	91	2	287	0.02	0.00	290.33	0.00	390.67
HAP2	14	91	3	328	0.03	0.00	333.53	0.01	423.57
STE12	13	78	4	131	0.05	0.00	139.93	0.00	282.90
RTG1	12	66	5	422	0.08	0.01	433.80	0.01	425.60
NDT80	11	55	2	610	0.04	0.02	613.53	0.01	352.37
RPN4	11	55	20	270	0.36	0.01	308.93	0.01	316.83
GCR2	11	55	3	367	0.05	0.01	372.07	0.00	281.03
ADR1	11	55	4	143	0.07	0.00	151.87	0.01	351.57
SKO1	11	55	0	89	0.00	0.00	90.00	0.00	143.07
SWI6	10	45	20	890	0.44	0.08	926.07	0.01	352.20
CAT8	10	45	31	448	0.69	0.03	510.73	0.01	353.07
MET4	10	45	27	441	0.60	0.02	496.40	0.00	354.53
HAA1	10	45	2	460	0.04	0.01	463.13	0.01	353.57
DAL81	10	45	11	349	0.24	0.01	371.40	0.01	359.80
PDR3	10	45	16	189	0.36	0.00	220.60	0.01	317.00
GAL4	9	36	10	246	0.28	0.01	265.60	0.01	283.90
GAT1	9	36	7	181	0.19	0.00	194.73	0.01	286.57
UPC2	9	36	3	101	0.08	0.00	150.80	0.01	213.40
RCS1	9	36	6	203	0.17	0.00	214.87	0.01	244.10
MAC1	9	36	5	65	0.14	0.00	75.00	0.01	281.13
HAP5	9	36	0	57	0.00	0.00	57.00	0.00	211.90
THI2	9	36	0	20	0.00	0.00	20.00	0.00	141.10
LEU3	8	28	28	459	1.00	0.03	516.13	0.01	283.53
VID30	8	28	6	406	0.21	0.02	416.87	0.00	286.17
HOG1	8	28	3	296	0.11	0.01	301.40	0.00	211.17
SWI4	8	28	7	333	0.25	0.01	346.47	0.01	281.90
SWI5	8	28	8	312	0.29	0.01	331.67	0.01	286.07
PHO4	8	28	1	148	0.04	0.00	150.73	0.00	175.83
TUP1	7	21	4	732	0.19	0.08	739.73	0.00	211.30
HAC1	7	21	1	188	0.05	0.01	191.73	0.00	174.50
HAP1	7	21	0	237	0.00	0.00	237.80	0.01	248.30
IME1	7	21	5	192	0.24	0.00	237.80	0.00	215.50
RIM101	7	21	0	47	0.00	0.00	46.93	0.01	178.17
MBP1	6	15	10	589	0.67	0.10	609.07	0.00	176.23
CYC8	6	15	7	526	0.47	0.06	539.33	0.01	175.70
SNF2	6	15	4	406	0.27	0.04	413.87	0.01	207.40
TYE7	6	15	1	246	0.07	0.01	247.67	0.01	173.00
SKN7	6	15	0	143	0.00	0.01	142.80	0.00	140.43
DAL82	6	15	7	186	0.47	0.00	199.87	0.00	215.60
RTG3	6	15	3	139	0.20	0.00	146.93	0.02	213.30
GRR1	5	10	8	459	0.80	0.04	474.13	0.01	175.77
RGT1	5	10	8	459	0.80	0.04	474.13	0.01	175.77
RFX1	5	10	1	391	0.10	0.03	394.33	0.00	139.50
SWI1	5	10	4	231	0.40	0.02	239.67	0.01	172.63
MET31	5	10	5	166	0.50	0.01	175.80	0.00	177.63
XBP1	5	10	0	165	0.00	0.01	164.87	0.01	181.80
HMRA1	5	10	0	91	0.00	0.01	90.87	0.00	103.13
SOK2	5	10	0	110	0.00	0.00	112.93	0.03	180.17
NRG1	5	10	1	102	0.10	0.00	105.93	0.02	178.00
RFA2	5	10	2	138	0.20	0.00	142.93	0.00	139.80
ACE2	5	10	5	220	0.50	0.00	233.00	0.02	181.47

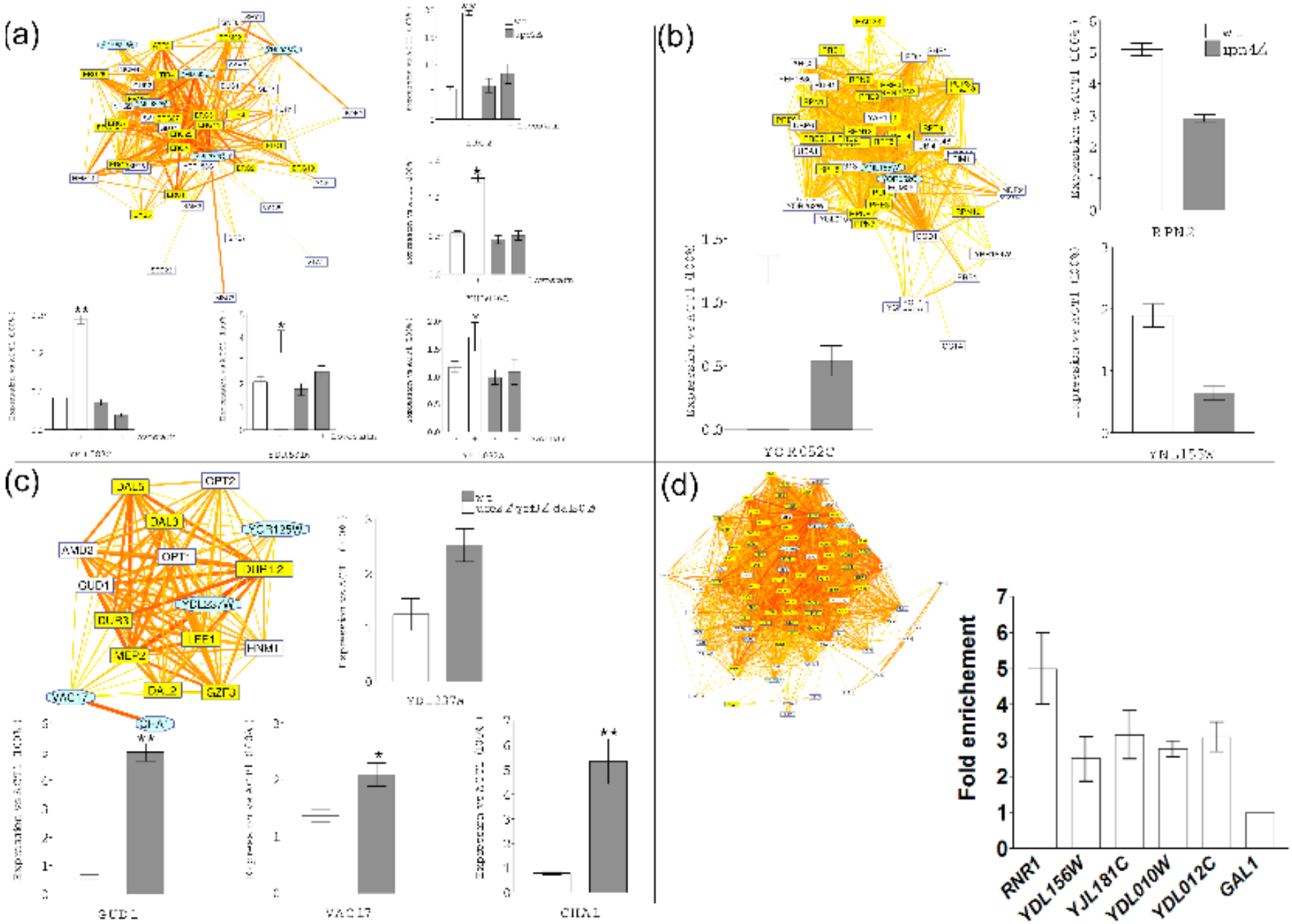
Inferred co-regulation versus annotated regulons



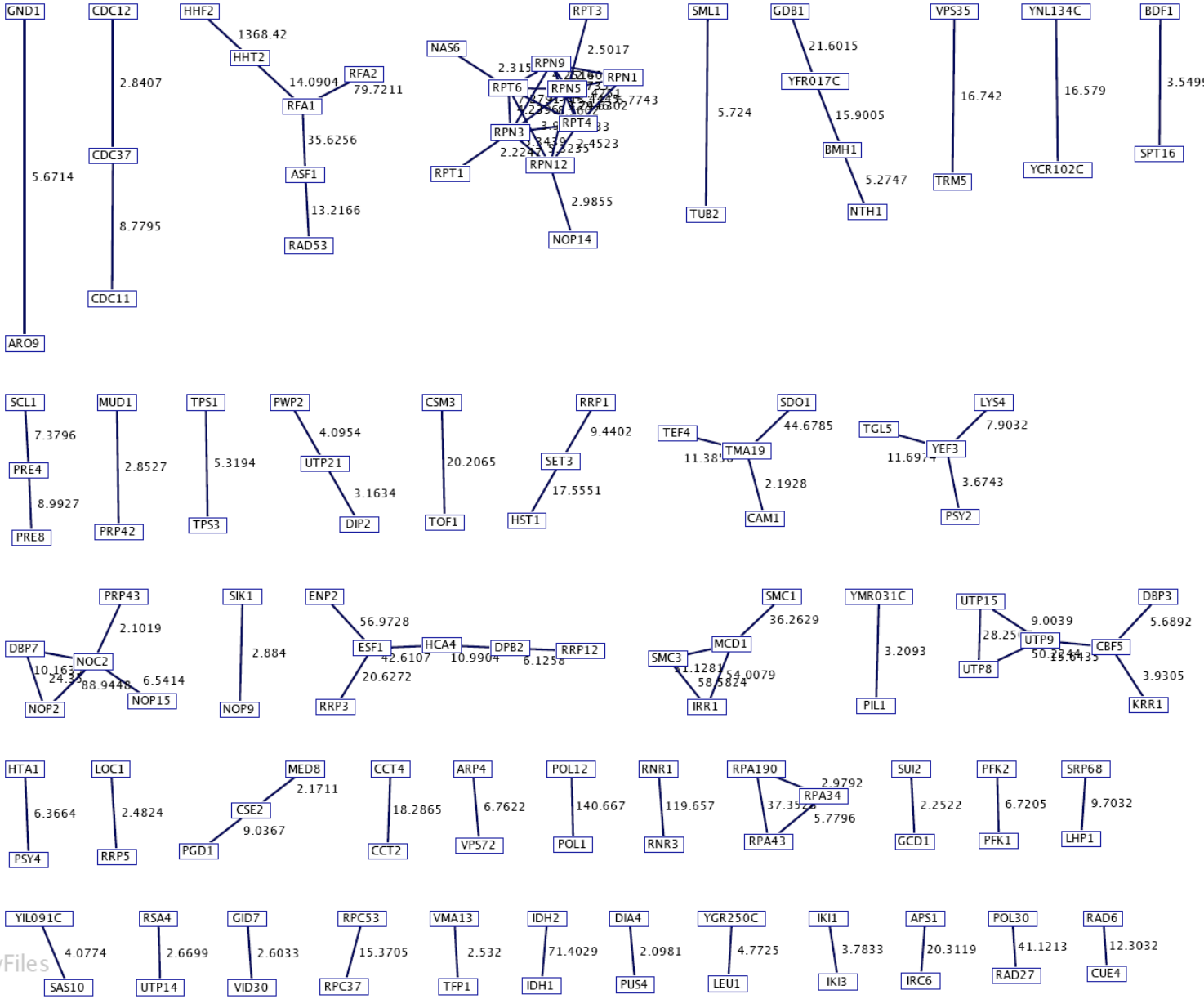
From phylogenetic footprints to co-regulation networks



Experimental validation of discovered motifs

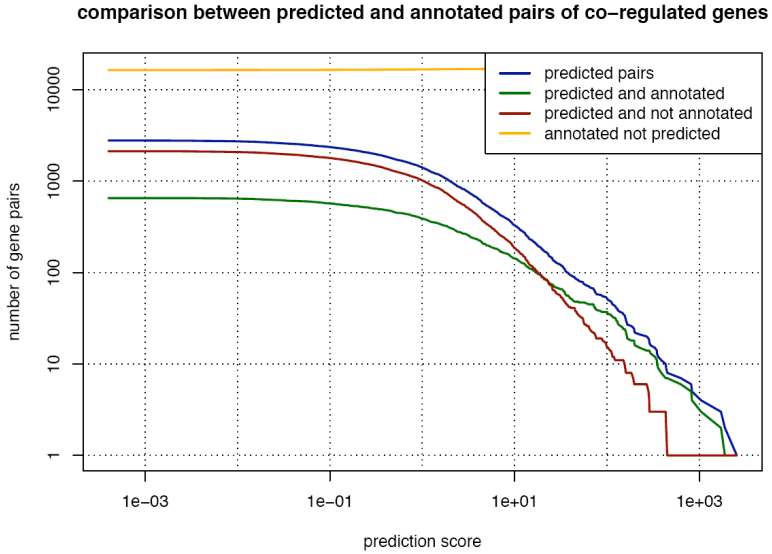


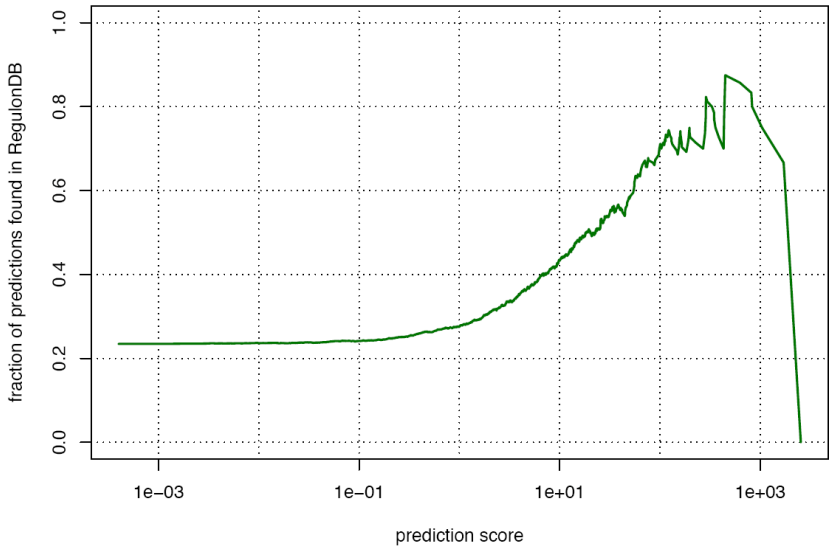
Inferred co-regulation versus protein interactions (Krogan 2006)



Supplementary material

Predicted versus annotated gene pairs



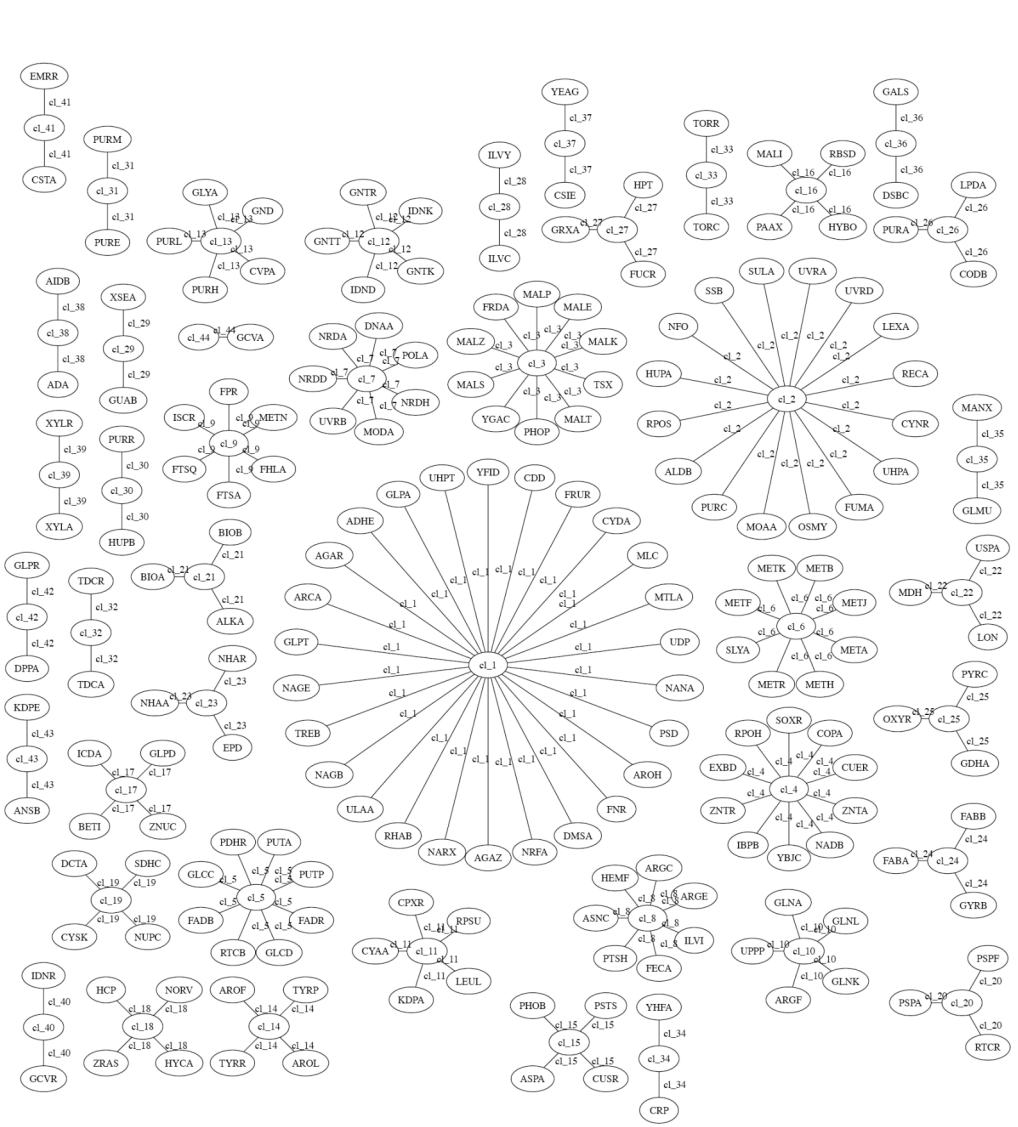


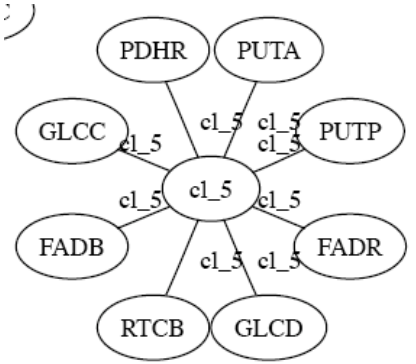
We performed a systematic estimation of the fraction of predictions found in RegulonDB, as a function of the prediction score (dot product).
More than 40% of the predictions with a score dp >= 10 are annotated in RegulonDB.
Note that predicted pairs which are not found in RegulonDB might nevertheless be correct, it is hard to estimate the rate of "false positive" in such a case.
The sensitivity can however be estimated, and it is quite low.

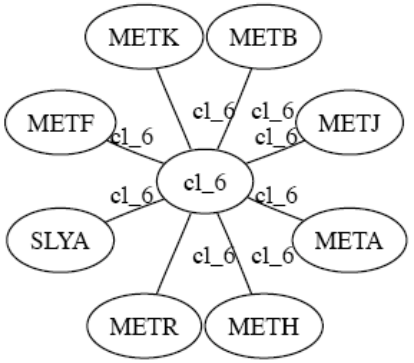
Among the 14,939 gene-gene co-regulated pairs in RegulonDB, the method detects no more than 652 (4.4%).
Note that more than 50% of the annotated gene pairs are linked via CRP.

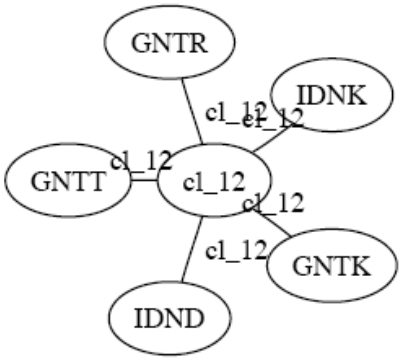
To do : analyze the sensitivity per TF, to check whether some regulons are preferentially detected.

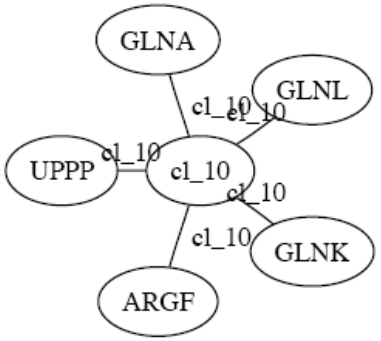
Clustering




PDHR
PUTA
GLCC
PUTP
FADB
FADR
RTCB
GLCD
cl_5


METK
METB
METF
METJ
SLYA
META
METR
METH
cl_6


GNTR
IDNK
GNTT
cl_12
GNTK
IDND

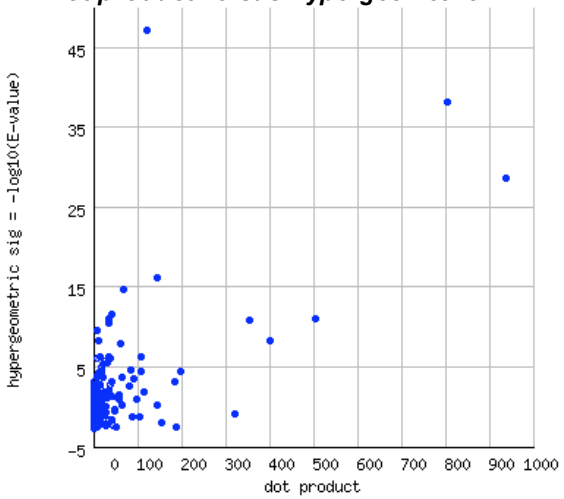

GLNA
GLNL
UPPP
cl_10
GLNK
ARGF

Clusters versus regulons

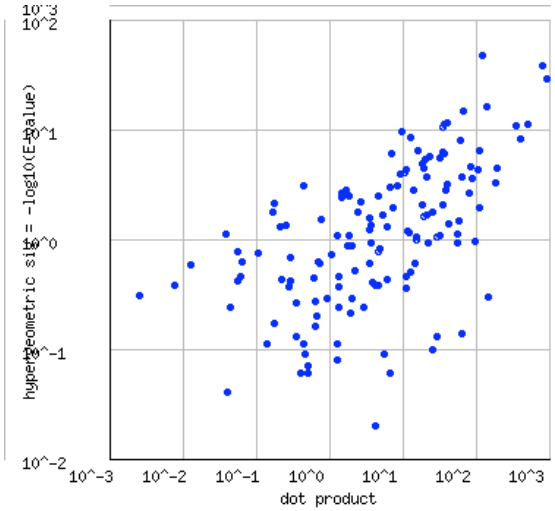
cluster	12	11	10	6	4	2	8	9	7	3	5	1	clust
CRP	22	4	10	1	6					4			47
FNR	4	1	1	7	2	3							18
IHF	5	2	1		1	2	2						13
ARCA	5	1	1	2									9
FIS	4			1		1				1		1	8
LEXA		1							1		6		8
NARL	4			1		3							8
PURR	7												7
FLHD	3				1	1		1					6
LRP	3	1									1		5
MALT			5										5
METJ								5					5
MODE	2			1		2							5
DNAA	1								2			1	4
MLC	2	1			1								4
NARP	3					1							4
TYRR							4						4
CPXR	2								1				3
CYTR	1		1							1			3
FRUR	1			1						1			3
GADX	2	1											3
GLPR	1				2								3
GNTR	3												3
MARA		1		1				1					3
NAGC		1			1					1			3
NTRC		3											3
PHOB							3						3
ARGP									1			1	2
BIRA							2						2
FADR	2												2
FHLA	2												2
GADE	1								1				2
GLCC		2											2
H-NS	1			1									2
IDNR	2												2
ILVY	2												2
MALI	2												2
METR	1							1					2
NAC	2												2

Compare-classes result (selection)

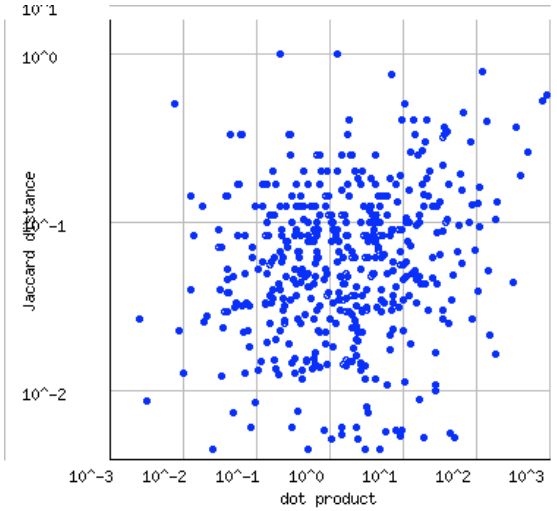
Dot product versus hypergeometric



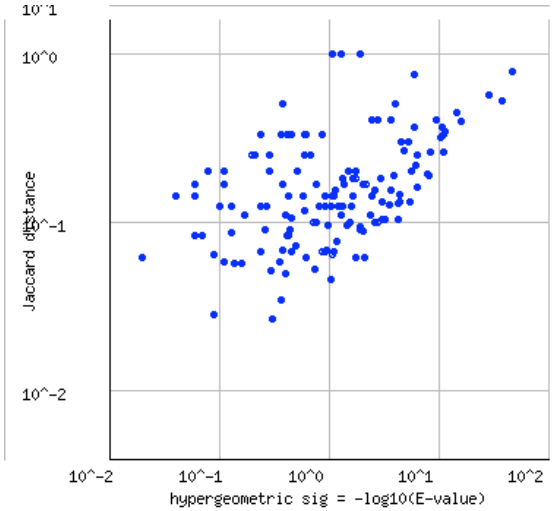
Dot product versus hypergeometric (log)



Dot product versus Jaccard (log)



Hypergeometricc versus Jaccard (log)



Discovered motifs in promoters of glpD orthologs from Enterobacteriales

Dyad	exp_freq	occ	exp_occ	occ_P	occ_E	sig	rank
atgn{1}tcg|cgan{1}cat	0.00069	11	1.02	1.30E-08	2.40E-04	3.62	1
aacn{0}att|aatn{0}gtt	0.00131	13	1.95	1.60E-07	3.00E-03	2.52	2
cgan{0}aca|tgtn{0}tcg	0.00056	9	0.84	2.90E-07	5.20E-03	2.28	3
agtn{10}tcg|cgan{10}act	0.00026	6	0.35	2.20E-06	3.90E-02	1.41	4
acan{13}acg|cgtn{13}tgt	0.00044	7	0.6	3.50E-06	6.40E-02	1.19	5
aacn{2}aca|tgtn{2}gtt	0.00082	9	1.21	5.60E-06	1.00E-01	0.99	6
aatn{2}tcg|cgan{2}att	0.00089	9	1.31	1.00E-05	1.90E-01	0.73	7
atgn{0}ttc|gaan{0}cat	0.00094	9	1.4	1.70E-05	3.10E-01	0.52	8
acgn{2}cat|atgn{2}cgt	0.00054	7	0.79	2.10E-05	3.80E-01	0.42	9
aacn{3}cat|atgn{3}gtt	0.00101	9	1.47	2.60E-05	4.60E-01	0.33	10
aacn{8}aag|cttn{8}gtt	0.00080	8	1.11	2.60E-05	4.70E-01	0.33	11
cctn{4}cac|gtgn{4}agg	0.00023	5	0.34	2.90E-05	5.30E-01	0.28	12
aacn{0}gaa|ttcn{0}gtt	0.00106	9	1.58	4.40E-05	7.90E-01	0.1	13
actn{7}aag|cttn{7}agt	0.00027	5	0.37	4.90E-05	8.90E-01	0.05	14

;assembly # 1	seed: atgntcg	13 words
aaangtt....	....aacnttt	0.82
.aatgtt....	....aacatt.	3.38
.aatnntcg..	..cgannatt.	2.1
..atgntcg..	..cgancat..	7.55
..atgttc...	...gaacat..	3.29
..atgnnngtt	aacnnncat..	1.63
..atgnncgt.	.acgnncat..	1.46
...tgttcg..	..cgaaca...	4.58
...tgtnngtt	aacnnaca...	2.4
...tgtncgt.	.acgnaca...	1.32
....gttngtt	aacnaac....	1.05
....gttcgt.	.acgaac....	0.47
.....ttcgtt	aacgaa.....	1.53
aaatgttcgtt	aacgaacattt	7.55
;assembly # 2	seed: atgnnnnnnntcg	19 words
aaangtt.........	.........aacnttt	0.82
.aatgtt.........	.........aacatt.	3.38
.aatnntcg.......	.......cgannatt.	2.1
..atgntcg.....	**.....cgancat..**	**7.55**
..atgttc........	........gaacat..	3.29
..atgnnnnnnntcg.	.cgannnnnnncat..	1.82
..atgnnngtt.....	.....aacnnncat..	1.63
..atgnncgt......	......acgnncat..	1.46
..atgnnnnnnnnnga	tcgnnnnnnnncat..	0.93
...tgttcg.......	.......cgaaca...	4.58
...tgtnngtt.....	.....aacnnaca...	2.4
...tgtncgt......	......acgnaca...	1.32
...tgtnnnnnntcg.	.cgannnnnnaca...	0.74
....gttngtt.....	.....aacnaac....	1.05
....gttcgt......	......acgaac....	0.47
....gttnnnnntcg.	.cgannnnnaac....	0.35
.....ttcgtt.....	.....aacgaa.....	1.53
......tcgnnntcg.	.cgannncga......	0.6
......tcgnnnncga	tcgnnnncga......	0.06
aaatgttcgttntcga	tcganaacgaacattt	7.55
;assembly # 3	seed: agtnnnnnnnnnntcg	3 words
agtnnnnnnnnnntcg	**cgannnnnnnnnnact**	**1.52**
.gttnnnnnnnnntcg	cgannnnnnnnnaac.	0.9
..ttcnnnnnnnntcg	cgannnnnnnngaa..	0.78
agttcnnnnnnnntcg	cgannnnnnnngaact	1.52
;assembly # 4	seed: acannnnnnnnnnnnnacg	3 words
acannnnnnnnnnnnnacg.	**.cgtnnnnnnnnnnnntgt**	**1.26**
acannnnnnnnnnnnnncga	tcgnnnnnnnnnnnnntgt	0.48
acannnnnnnnnnnnaac..	..gttnnnnnnnnnnntgt	0.01
acannnnnnnnnnnnaacga	tcgttnnnnnnnnnnntgt	1.26

Selected inferred clusters

id	name	description [synonyms]
YGL255W	ZRT1	High-affinity zinc transporter of the plasma membrane, responsible for the majority of zinc uptake; transcription is induced under low-zinc conditions by the Zap1p transcription factor [YGL255W;ZRT1;ZRT1;YGL255W;852637;6321182;NP_011259]
YKL175W	ZRT3	Vacuolar membrane zinc transporter, transports zinc from storage in the vacuole to the cytoplasm when needed; transcription is induced under conditions of zinc deficiency [YKL175W;ZRT3;ZRT3;YKL175W;853679;6322673;NP_012746]
YLR130C	ZRT2	Low-affinity zinc transporter of the plasma membrane; transcription is induced under low-zinc conditions by the Zap1p transcription factor [YLR130C;ZRT2;ZRT2;YLR130C;850821;6323159;NP_013231]
YOL154W	ZPS1	Putative GPI-anchored protein; transcription is induced under low-zinc conditions, as mediated by the Zap1p transcription factor, and at alkaline pH [YOL154W;ZPS1;ZPS1;YOL154W;854011;6324419;NP_014488]
YGL258W	VEL1	Protein of unknown function; highly induced in zinc-depleted conditions and has increased expression in NAP1 deletion mutants [YGL258W;VEL1;VEL1;YGL258W;852634;6321179;NP_011256]
YNR040W	YNR040W	Putative protein of unknown function; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies [YNR040W;YNR040W;855776;6324368;NP_014438]

Phylogenetic footprints
MET genes in Gammaproteobacteria