DISTILLER: query-driven application

Goal:

Starting from a set of genes of interest (referred to as query genes), find out which genes are co-expressed with these query genes, the conditions in which these genes are co-expressed and the regulatory motifs that are present in the promoter region of these genes.

Information:

Query genes:

- *rne, eno, pnp, rhlB, rnb, rnc, rng, pcnB, rnr, rraA, rraB, gmr, csdA*

A priori Information on the regulation:

- *pcnB* may be more highly expressed during stationary phase
- *eno* is positively regulated by Cra
- *rraB* could be regulated by ArgR
- *rraA* could be transcribed by sigmaS

Data sources:

Expression data:

- the expression compendium from Lemmens et al., Genome Biology, 2009.

Motif data:

The 67 motif matrices of RegulonDB were used to screen the intergenic regions of *E. coli*. We combined two different screening strategies to find as many motif hits as possible:

- The first screening strategy = Hertzberg et al., 2005
- The second screening strategy = Clover (Frith et al., 2004)

Detection of transcriptional modules:

DISTILLER integrates several data sources to detect sets of co-expressed genes, the conditions in which they are co-expressed and the motifs present in the promoter region of the co-expressed genes. DISTILLER was applied on the expression and motif data. Since DISTILLER uses an item set mining strategy, all possible solutions (i.e. all significant modules) will be in the output.

In Lemmens et al. (2009) the most interesting modules are selected from this total list of modules in a greedy way.

In this particular application, our goal is to obtain modules that consist of at least one query gene. Therefore, we started from the initial output and searched for those modules containing one or more query genes. Instead of applying our greedy post-processing step, we kept all modules (obtained via runs of different DISTILLER parameter settings) containing query genes as they might all be of interest to the user. As a result, many modules that still overlap in their gene and condition content will be selected.

In table 1 the number of modules that were identified for each query gene are shown. For instance, when we use gene *rne* as a query gene, DISTILLER identified 4 modules that consist of co-expressed genes (including *rne*). All of these modules are under control of TorR, meaning that the genes in each module have a motif instance for TorR in their promoter region. We also checked whether the genes in the 4 modules were already known to be regulated by TorR, but this was not the case for these modules (No. RegulonDB hits = 0).

Table 1: Summary of the results obtained for each query gene. The total number of modules for each query gene, the number of different motifs contained by the modules, the number of query genes and the number of hits in RegulonDB are shown.

Query_gene	No. modules	No. motifs	Motifs	No. query genes	Query genes	No. RegulonDB hits
b1084 (rne)	4	1	TorR	1	b1084	0
b2779 (eno)	0	0		0		0
b3164 (pnp)	20	1	Ada	1	b3164	0
b3780 (rhlB)	349	1	ArcA	1	b3780	22
b1286 (rnb)	0	0		0		0
b2567 (rnc)	7	1	UxuR	1	b2567	0
b3247 (rng)	28	1	Mlc	1	b3247	1
b0143 (pcnB)	13	2	GcvA UlaR	1	b0143	1
b4179 (rnr)	209	1	Fur	1	b4179	21
b3929 (rraA)	424	4	ArgP DcuR FNR LexA	1	b3929	47
b4255 (rraB)	109	3	ArgR IscR NsrR	1	b4255	9
b1285 (gmr)	668	3	ArcA Fur NsrR	1	b1285	30
b2810 (csdA)	14	2	FlhDC RcsB	1	b2810	6

Post-processing of the modules

To find out the genes and conditions that are represented most frequently in the modules (and that thus correspond to the most pronounced signal in the data), modules were grouped together in a cluster if the same query gene was present in the module and the genes in the module were regulated by the same regulator. For each group of modules, genes were ranked according to their presence in those modules. The gene that is present in most modules is ranked the highest, whereas the gene that is present in the smallest number of modules, is ranked the lowest. The same strategy was applied on the conditions.

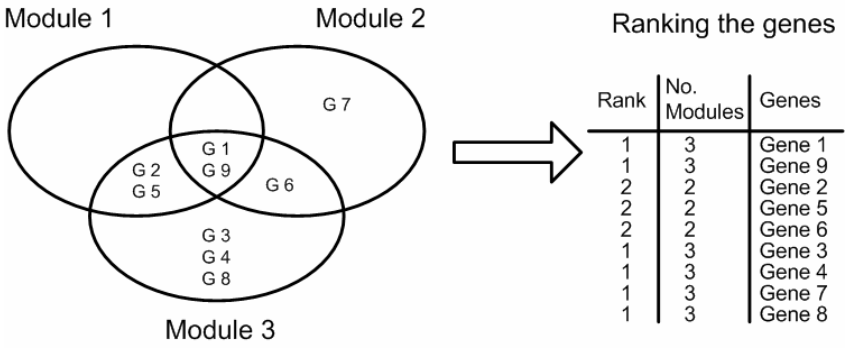

Figure 1: Ranking the genes. Modules are grouped together if the same query gene was present in the module and if the genes in the modules have a motif instance for the same motif. The gene that is present in most modules is ranked the highest, such as Gene 1 or Gene 9.

Table 2: Summary of the results per query gene. The number of modules that contains the query gene, the top 5 most frequent genes and the corresponding motif are shown.

Query gene	Number of modules	Most frequent genes	Motif
b1084 (rne)	4	b1084 b0754 b1782 b1452 b2495	TorR
b3164 (pnp)	20	b3164 b0274 b0021 b0264 b0988	Ada
b3780 (rhlB)	349	b3780 b0432 b0431 b0126 b0430	ArcA
b2567 (rnc)	7	b2567 b1277 b3165 b2329 b2794	UxuR
b3247 (rng)	28	b3247 b3168 b0167 b2974 b3147	Mlc
b0143 (pcnB)	6	b0143 b1186 b2113 b0413 b0125	UlaR
	7	b0143 b1281 b1277 b2905 b3002	GcvA
b4179 (rnr)	209	b4179 b4367 b0150 b1452 b0584	Fur
b3929 (rraA)	371	b3929 b1004 b0919 b1003 b2844	FNR
	21	b3929 b1004 b2009 b0209 b1003	DcuR
	2	b4014 b3929 b1004 b2408 b2844	ArgP
	30	b3929 b2699 b0231 b0140 b1848	LexA
b4255 (rraB)	19	b4255 b1540 b0741 b2145 b1289	IscR
	62	b4255 b3172 b4254 b3958 b3957	ArgR
	28	b4255 b1259 b1428 b3094 b1327	NsrR

b1285 (gmr)	477	b1285 b3006 b4367 b1452 b0150	Fur
	53	b1285 b0431 b1415 b0429 b0430	ArcA
	138	b1285 b1428 b0485 b2097 b1259	NsrR
b2810 (csdA)	6	b2810 b1550 b0957 b2038 b2007	RcsB
	8	b2810 b1940 b1944 b1938 b1074	FlhDC

A more detailed description of the resulting modules per query gene can be found in the HTML file. Below we explain what is shown in the HTML file by means of an example: The query gene b4255 (*rraB*) that was present in 62 modules regulated by ArgR.

The motif logo derived from RegulonDB is shown in the HTML file.

The genes and conditions for these 62 modules were ranked based on their presence in the modules. For instance, if a gene is present in all modules, this gene was ranked the highest. The "gene information" table in the HTML file shows the top 10 of the highest ranked genes. By clicking on "View all genes", a list of all ranked genes is shown, together with the number of modules in which they appear. If a gene is one of the query genes, this is indicated in the column "Query Gene". If the gene is known to have an instance of the motif in its upstream region according to RegulonDB, this is indicated in the column "RegulonDB". The best scoring motif instances according to the method of Hertzberg et al. or Clover are shown in the table as well.

The "condition information" table in the HTML shows the 20 highest ranked conditions. By clicking on "View all conditions" a list of all ranked conditions is shown, together with the number of modules in which they appear. The conditional class to which the array belongs is shown in the column "Conditional class". This conditional class indicates the major cue that was measured on the microarray. The genes that are in modules regulated by regulator ArgR, for instance, are often co-expressed in conditions measuring the effect of addition or depletion of amino acids.

Genes and conditions that are ranked the highest appear thus in most modules. We can therefore say that we are most confident that these highly ranked genes and conditions are truly associated with the query gene and may be the most interesting for further investigation. If we have a look at all ranked conditions for our example, we see that none of the conditions was present in all modules, but the highest ranked condition, for instance, was present in 54 out of 62 modules. This indicates that in these highest ranked condition, genes of 54 modules will be co-expressed but the

genes of 8 modules will not be co-expressed in this condition. The latter can also be concluded when checking the 6 plots in the HTML file.

The first 3 plots show the heatmap and expression profile (only the 50 highest ranked conditions) for all genes, while the last 3 plots show the heatmap and expression profile for the 10 highest ranked genes.

When we have a look at the heatmap or the expression profile for the 10 highest ranked genes, it is clear that almost all of these genes are co-expressed in the best ranked conditions. Some genes like *ygeK* however show a different behaviour in these conditions. This is an indication that these genes belong to a module for which the genes are not co-expressed in this condition set. Nevertheless there were 54 other modules for which this condition is indeed important.

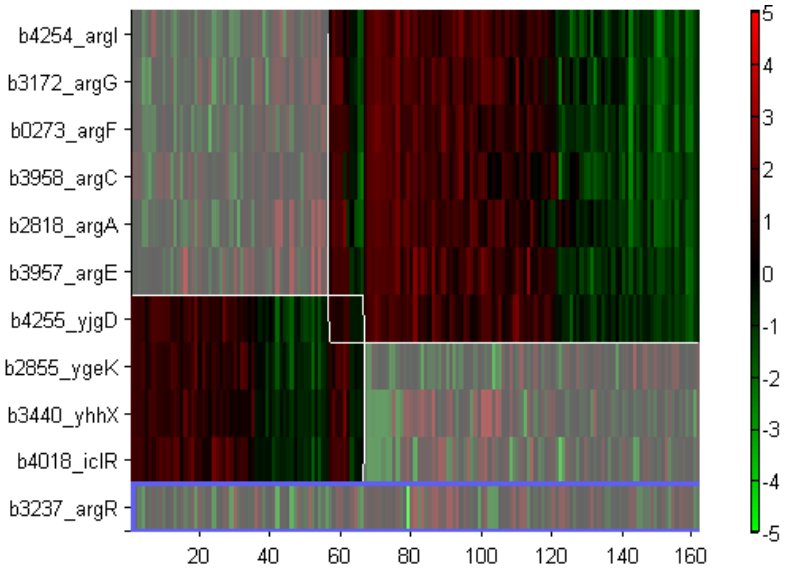

Figure 2: The heatmap of the expression profiles for two modules are shown. The first module consists of the genes *argI*, *argG*, *argF*, *argC*, *argA*, *argE*, and *yjgD* (= *rraB).* The second module consists of *yjgD* (= *rraB*), *ygeK*, *yhhX*, and *iclR*. These modules show only a small overlap in their condition content (from condition 55 until condition 65). Condition 76 is the highest ranked condition over all modules and present in 54 out of 62 modules. The second module however does not contain this condition, and thus the genes from this module are not expected to be co-expressed in this condition.

Figure 2 shows two of the modules that were identified by DISTILLER. The first module consists of the genes *argI*, *argG*, *argF*, *argC*, *argA*, *argE*, and *yjgD* (= *rraB).* The second module consists of *yjgD* (= *rraB*), *ygeK*, *yhhX*, and *iclR*. Modules are indicated by grey boxes. Condition number 76 is the highest ranked condition among all modules. This condition is indeed present in the first module, but not in the second. The genes of the second module are thus not co-expressed in the highest ranked condition.