Application of Regulatory Sequence Analysis and Metabolic Network Analysis to the Interpretation of Gene Expression Data

Jacques van Helden[1,2], David Gilbert[2,3], Lorenz Wernisch[2] , Michael Schroeder[3], and Shoshana Wodak[1,2]

[1] SCMBB. Université Libre de Bruxelles
CP160/16. 50 av F.D. Roosevelt. B-1050 Bruxelles. Belgique.
{jvanheld,shosh}@ucmb.ulb.ac.be
[2] European Bioinformatics Institute.
Genome Campus - Hinxton Cambridge CB10 1SD - UK.
{jvanheld,drg,lorenz,shosh}@ebi.ac.uk
[3] Department of Computing, City University.
Northampton Square, London EC1V 0HB, UK.
{drg,msch}@soi.city.ac.uk

We present two complementary approaches for the interpretation of clusters of co-regulated genes, such as those obtained from DNA chips and related methods. Starting from a cluster of genes with similar expression profiles, two basic questions can be asked:
1. Which mechanism is responsible for the coordinated transcriptional response of the genes? This question is approached by extracting motifs that are shared between the upstream sequences of these genes. The motifs extracted are putative cis-acting regulatory elements.
2. What is the physiological meaning for the cell to express together these genes? One way to answer the question is to search for potential metabolic pathways that could be catalyzed by the products of the genes. This can be done by selecting the genes from the cluster that code for enzymes, and trying to assemble the catalyzed reactions to form metabolic pathways.
We present tools to answer these two questions, and we illustrate their use with selected examples in the yeast *Saccharomyces cerevisiae*. The tools are available on the web (http://ucmb.ulb.ac.be/bioinformatics/rsa-tools/; http://www.ebi.ac.uk/research/pfbp/; http://www.soi.city.ac.uk/~msch/).

1 Introduction

DNA chips (2-4) and related techniques permit the measurement of the transcriptional response of all the genes of an organism to a controlled stimulus (presence/absence of metabolites, action of a drug, temperature, …) or to a genetic modification (deletion or over-expression of a selected gene). Results of several experiments are combined into a multivariate table, summarizing the response of all the genes of

an organism to a variety of conditions. Genes can then be clustered on the basis of similarities in their expression profiles. Different approaches have been used for this purpose: hierarchical clustering (4), self-organizing maps (17), k-means (25). Once such clusters have been obtained, two complementary questions can be asked (Figure 1).

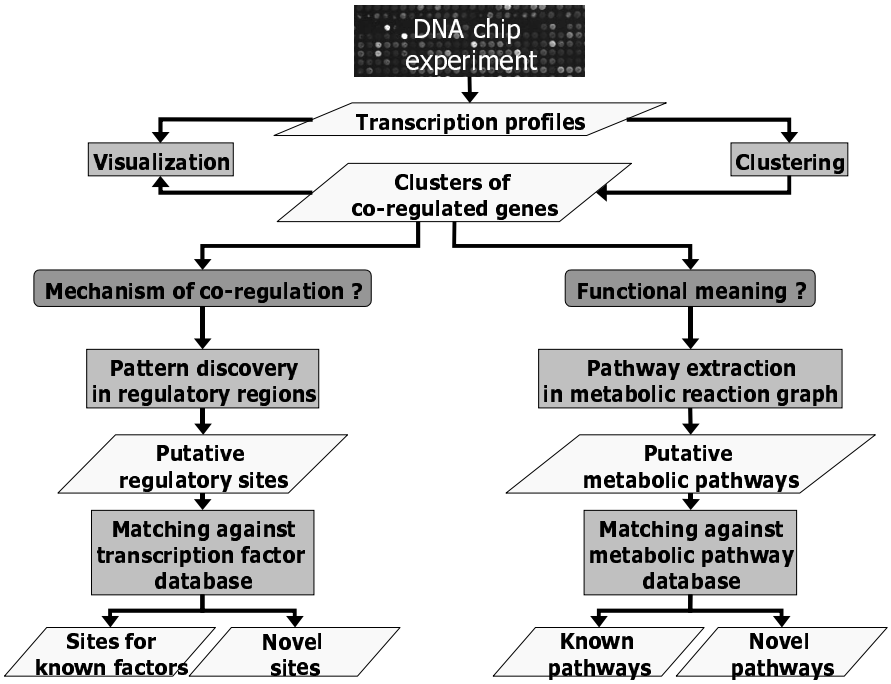

Figure 1: Flow chart of the approaches for interpreting gene expression data. Round shaded rectangles represent the questions, shaded rectangles the processes, and white parallelograms data.

(1) Which mechanism could ensure the coordinated transcriptional response of the genes belonging to a given cluster? Transcriptional regulation is mediated by a class of proteins, called transcription factors, which bind specific DNA motifs, called cis-acting elements, and interact with RNA polymerase to enhance (activation) or reduce (repression) the expression of neighboring genes. The answer to this question amounts to a search for transcription factors that might act simultaneously on the genes belonging to the cluster. There is no obvious way to predict directly which proteins act in trans on a set of genes, but the problem can be addressed indirectly by first predicting which cis-acting elements might be involved, and then looking for transcription factors that might bind these cis-acting elements. The approach consists in analyzing upstream sequences to discover shared motifs, which could correspond to regulatory elements. Candidate cis-acting elements can then be matched against databases of known binding sites (15, 26), and/or tested experimentally.

(2) Which biological function requires a coordinated expression of the genes belonging to the cluster under consideration? It is usual that genes involved in a common process are co-regulated, ensuring the presence of all the necessary proteins. The question amounts thus to search for processes in which most genes of the cluster might be involved. One simple approach is to match the set of genes against a database of gene/protein function (23). This would however restrict the possible answers to processes/pathways that have already been previously characterized and are stored in the database. A more flexible approach is to try to identify reactions that could be catalyzed by the gene products, and to interconnect these reactions in any possible way to generate potential metabolic pathways. The pathways assembled by this way can then be matched against metabolic pathway databases. Part of these pathways will correspond to previously described pathways, whereas in other cases one should be able to discover novel pathways.

This paper is a mini-review of our recent work on several aspects related to the interpretation of gene expression data. We illustrate the different questions that can be addressed on the basis of a selected study case, and discuss some critical issues for obtaining suitable results. We refer to previously published work for a detailed description of the statistical and algorithmic aspects, which would go beyond the scope of this review.

2 Gene Expression Data : a study case

To illustrate our purpose, we selected an example of gene expression data from the literature. Spellman and co-workers used the DNA chip technology to detect yeast genes that are involved in cell cycle (16). These authors measured the level of expression of all 6000 yeast genes at different time points during the cell cycle, and selected those showing periodic fluctuations. The 800 selected genes were then clustered according to similarities in their expression profiles. Some of the clusters obtained were clearly associated to well defined cellular processes associated to the cell cycle. An unexpected cluster was also isolated, mostly made of genes involved in methionine biosynthesis. We will use this MET cluster as study case throughout the following chapters.

3 Visualization

The development of flexible and intuitive visualization tools is an important requirement for the interpretation of gene expression data (Figure 1). One popular approach has been to apply hierarchical clustering and to display the profiles of expression in parallel with the dendrogram (5). We are currently working on complementary approaches, which would provide a direct representation of the functional distances between genes (7). This is illustrated in Figure 2, which shows a mapping of the 800 genes from Spellman's experiment on a Euclidian space. Each dot represents a single

gene. Coordinates were assigned so that the distances between dots reflect the dissimilarities between gene expression profiles. Noticeably, genes are grossly aligned along a ring, which is probably the most direct way to represent cell cycle. In particular, the center of the ring is avoided, and most genes align on the periphery, whereas random data would occupy the center as well as the periphery (not shown). Genes with synchronous fluctuations of expression appear in the same angle of the circle. The MET cluster mentioned above is highlighted.

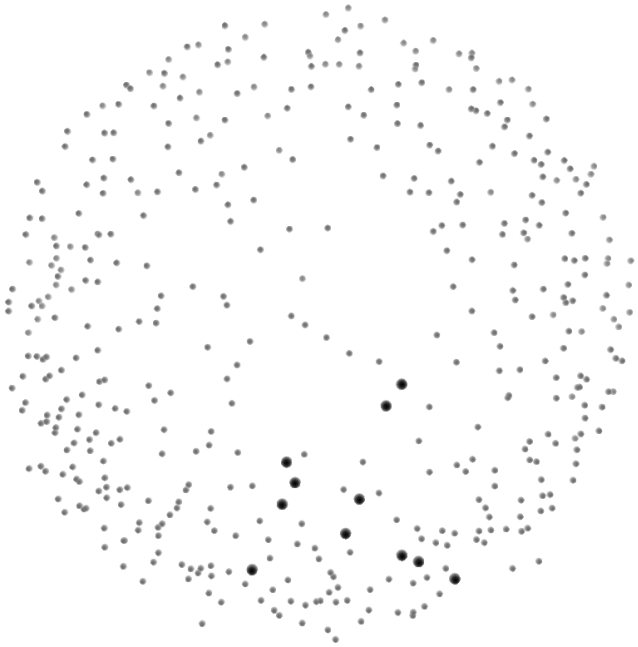

Figure 2: Visualization of gene expression data. Spellman and co-workers (16) used DNA chips to measure the level of expression of all yeast genes during cell cycle, and isolated 800 gene showing periodical fluctuations. In the original paper, genes are displayed on a tree. In the alternative representation shown above, genes are mapped on an Euclidian space. Each dot represents a gene. Coordinates are assigned so that the distance between two dots reflects the distance between the corresponding gene expression profiles. In the case of cell-cycle regulated genes, most dots align on the periphery of a ring, and the center is avoided (which would not be expected from random data). Genes having a synchronized peak of expression are located in the same angle of the ring. The genes from the MET family, our study case, are highlighted. Note that the visualization programs support true colors, and would allow to discriminate several clusters on the same image (7). The present image is in grayscale due to publication restrictions.

4 Regulatory Sequence Analysis

Features of Cis-Regulatory Elements

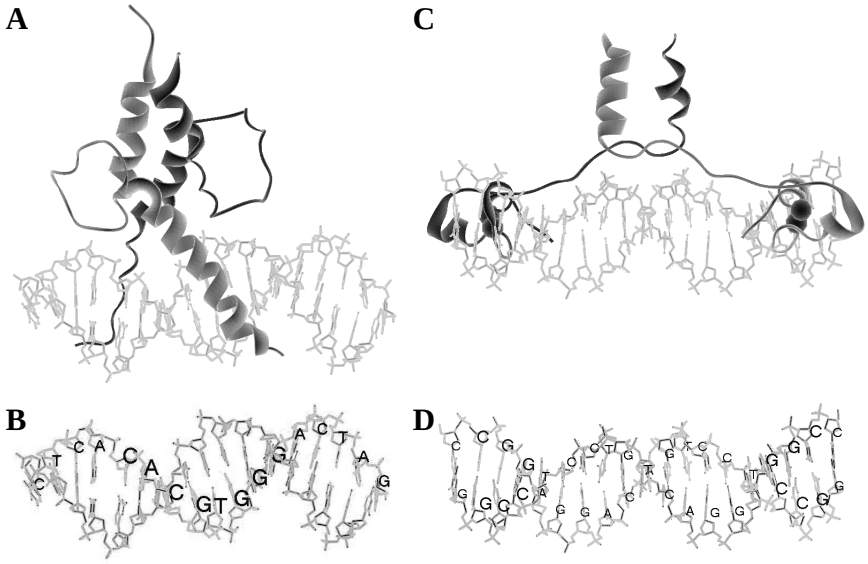

Figure 3: Structural features of transcription factor-DNA interfaces. **A**. Pho4p-DNA complex. Notice the two protein monomers acting on DNA like tweezers, and the restricted number of adjacent nucleotides that enter into direct contact with the protein. **B.** Pho4p binding site. The central oligonucleotide shown in larger characters (CACGTGGG) is conserved among Pho4p between sites. **C.** Gal4p-DNA complex. Gal4p belongs to the Zn cluster family of transcription factors. Notice the pair of trinucleotides entering into direct contact with the protein, separated by an intermediate region. **D.** Gal4p binding site. In contrast with Pho4p, the conserved region is not an oligonucleotide but a pair of trinucleotides (GGC-GGC) separated by a 11bp region of weakly conserved bases.

Transcription factors bind to short stretches of DNA, whose sequence is highly specific for each particular transcription factor. The specificity of the binding site is determined by the structure of the protein domain that enters into direct contact with DNA. Most sites described can be regrouped into two classes. The first class consists of a short sequence of highly conserved adjacent nucleotides (typically 6 base pairs), surrounded by a few partly conserved nucleotides (Figure 3A,B). This kind of site is associated to proteins containing different types of DNA-binding domains: Zn finger, homeodomain, leucine zipper, basic Helix-Loop-Helix. Another class of binding sites consists of a pair of very short oligonucleotides (typically 3 base pairs each), separated by a region of fixed width but variable content (Figure 3C,D). These sites are

typical for factors containing a Zn cluster domain (found in fungi) or a Helix-Turn-Helix domain (the most frequent DNA-binding domain in prokaryotes). Many of these sites show an internal symmetry (tandem repeat or reverse palindrome) due to the fact that the transcription factor binds DNA in a homodimeric form, each monomer entering into direct contact with one trinucleotide.

In the yeast *Saccharomyces cerevisiae*, cis-acting elements are found upstream the genes they control, within a range of 800 base pairs from the start codon. Their efficiency generally depends neither on their precise location nor on their strand orientation. It is common to find several sites for the same transcription factor within a same upstream region. These repetitions allow a synergic action of several copies of the protein.

String-based Approaches to Pattern Discovery in Regulatory Sequences

Based on these properties of yeast cis-acting elements, we developed two specialized programs for discovering putative regulatory elements within a set of upstream regions.

The first program, *oligo-analysis*, performs a statistical analysis of single word frequencies (typically hexanucleotides) in a set of sequences, and extracts all words which are significantly over-represented. In a previous paper (19) we showed that, despite its simplicity, this program is able to extract many regulatory sites with a very low rate of false positives.

The single-word analysis however fails to detect some cis-acting elements, especially those bound by Zn cluster proteins. This is not surprising, since in these sites the conserved nucleotides are shared between two sub-sites separated by a variable region. Consequently, we developed a complementary program, *dyad-analysis* (24), which performs a statistical analysis of all possible pairs of shorter words (typically trinucleotides) with different spacing values (between 0 and 20). This program is very efficient, not only for the detection of sites bound by Zn cluster and HTH proteins, but also for sites recognized by other classes of transcriptional factors.

Criteria for the Statistical Analysis of Words and Dyads

The efficiency of the above mentioned programs crucially depend on the choice of appropriate parameters.

Selection of non-redundant upstream sequences

Before envisaging any analysis of upstream sequences, it is essential to avoid redundancy in the data set. Indeed, the applicability of the statistical tests relies on the mutual independency of the sequences. Two sources of redundancy can be identified.

- Recent duplication of a gene, together with its upstream region. Such duplications are particularly common in yeast telomeric regions.

- Intergenic region shared between two divergently transcribed neighbor genes. If both genes belong to the same cluster, the intermediate upstream region will be included twice in the sequence set.

The data set has thus to be purged by discarding sequences that show a high similarity with the direct or reverse strand of any other sequence of the set.

How to count word occurrences?

Should words be counted on a single or both strands? How to treat overlapping occurrences of a same word? The choice of the counting mode depends on the expected characteristics of regulatory sites, which can differ between organisms. In the case of yeast, best results are obtained by counting occurrences on both strands and without overlap.

Estimation of Expected Word Frequencies

The simplest approach would be to consider all words as equiprobable. This would however provide bad results, due to the high frequency of A and T nucleotides in yeast sequences. This bias can be corrected by calculating expected word frequencies on the basis of nucleotide-specific frequencies (residue frequencies). This correction already provides better results, but still returns many false signals, mainly AT-rich sequences, due to a preferential aggregation of A and T nucleotides in yeast non-coding sequences. The best approach consists in calculating pre-defined tables of expected word frequencies (background frequencies), based on the whole set of yeast non-coding sequences.

How to Compare Expected and Observed Frequencies?

Several statistics have been envisaged for detecting over-represented words in DNA sequences: observed/expected ratio (1), log likelihood (8), Poisson distribution (14), Z-values (21), binomial distribution (19, 24). The observed/expected ratio has to be avoided, because it is strongly biased in favour of patterns with low expected frequencies. The log-likelihood introduces a correction for this bias, but is not easy to convert to probabilistic values. Z-scores rely on an assumption of normality of the distribution of occurrences, which is only verified when sequence length tends towards infinite. For small sequence sets (a few thousands of base pairs) such as families of upstream sequences from co-regulated genes, the most appropriate statistics are Poisson and binomial.

How to Select the Threshold of Probability?

Analyzing a single sequence set involves a comparison between observed and expected frequencies for several thousands of words. For example, there are 4,096 possible hexanucleotides. This means that with a probability threshold of 0.01, around 40 words would still be selected from any random sequence. According to the Bonferoni rule, these false positive can be avoided by lowering the threshold of probability to a value lower than 1/4096=0.00025. For heptanucleotides, the threshold should be

lowered to 1/16,564. Thus, the threshold value has to be adapted to the number of words taken into consideration, which itself depends on the word length.

Significance index

We defined a significance index (19, 24) which provides an intuitive way to evaluate the degree of over-representation.

$$sig = -\log_{10}\left[P(occ \geq n) * D\right]$$

Where *D* is the number of possible patterns, and *P(occ≥n)* the probability to observe n or more occurrences for the word considered. This probability can be calculated with the binomial, Poisson or normal forumula, as described above.

The significance index takes into consideration the effects mentioned above, including the reduction of significance when the number of possible patterns increase. The index can take positive or negative values. The interpretation is fairly intuitive. Positive values indicate over-representation. In random sequence sets, one expects to find no more than one pattern with a positive value, independently of the conditions such as sequence length, number of sequences, pattern length, ... A value higher than 1 is expected every 10 sequence sets, a value higher than 2 every 100 sequence sets, and more generally a value higher than s is expected every 10^s sequence sets. The index applies to spaced dyads (24) as well as single oligonucleotides (19).

How to Select Word Length?

Small words (di- to tetra-nucleotides) present a marked bias from theoretical distributions, and Poisson/binomial statistics are thus inappropriate. On the other side, analyzing too large words (octa-, nona-nucleotides) would prevent any of them to be detected as significant. Practically, we observed that hexanucleotide analysis provides excellent results in most cases for yeast sequences (19).

Even when restricting the analysis to hexanucleotides, larger patterns can nevertheless be detected, by assembling strongly overlapping hexanucleotides.

All the statistical considerations above are easily extended to the analysis of spaced dyads (24).

Evaluation of String-based Approaches with Known Regulons

In order to evaluate the above methods, sets of genes were collected for which the transcription factor was already known. The programs were fed with the upstream sequences, and the significant patterns were compared with the expected binding sites (19, 24). Generally, the number of significant words/dyads is restricted to a dozen per gene set. Some of these selected words/dyads strongly overlap with each other, and can be combined (using a custom fragment assembly program called *pattern-assembly*) to form a larger pattern. Pattern assembly also allows to describe, to some extent, the partial degeneracy of some binding sites. Indeed, it is frequent to detect several patterns that differ by a single substitution, and correspond to variants recoginzed by the same transcription factor. Patterns with a high significance index are

always associated with known transcription factors. Some additional patterns appear that might be associated with novel transcription factors.

Application to Clusters Obtained from Microarray and Related Technologies

After having evaluated the programs with known regulons, we applied the same string-based approaches to extract putative regulatory elements from clusters of genes resulting from DNA chip experiments. We published elsewhere (24) an analysis of families of cell cycle regulated genes defined by Spellman (16). We show here in more details the results obtained with the MET family (Table 1). All pairs of trinucleotides separated by spacing between 0 and 20 were anaylzed. The significant patterns form three groups of overlapping words, that can be assembled into 3 larger patterns. One additional isolated dyad is selected.

The first group of words corresponds to the binding site of the Met4p/Met28p/Cbf1p complex. The second group corresponds to Met31p and Met32p binding sites. All these transcription factors are known to act cooperatively to activate transcription of genes related to methionine metabolism. The highest score within each group is highlighted in bold. The pattern selected with the highest score generally corresponds to nucleotides that enter into direct contact with the transcription factor.

It is not possible to evaluate the efficiency of the programs on families obtained from DNA chip experiments with the same precision as was done with known regulons, since the transcription factors are usually not known. However, we observed that the same kind of result is generally obtained: a very restricted number of words/dyads are selected as significant. For some families, patterns are selected with a very high significance index, suggesting a very likely putative regulatory element. In other families, the patterns selected have a lower significance index. This is often the case for very small families (less than 5 genes), and results from a reduction of the signal-to-noise ratio. On the other extreme, analyzing too large gene clusters (> 50) reduces the sensitivity of the programs. The reason is that the larger clusters are less likely to be regulated by a single factor, and might contain a mixture of different signals. The effect of mixing together sequence that contains a given signal with sequences that do not contain it is also to reduce the signal-to-noise ratio. The programs are able, to some extent, to extract multiple signals from a single analysis, but the highest efficiency is clearly obtained by selecting clusters of genes that are likely to be all regulated by the same transcription factor. The choice of the clustering method is thus crucial.

	pattern	reverse complementary	obs occ	exp occ	proba	sig
group 1	GTC..GTG..	..CAC..GAC	17	2.61	3.60E-09	3.8
	.TCA.GTG..	..CAC.TGA.	23	5.12	8.50E-09	3.4
	.TCACGT...	...ACGTGA.	21	4.75	4.60E-08	2.7
	..CACGTG..	**..CACGTG..**	**38**	**3.37**	**0**	**20**
	..CAC.TGA.	.TCA.GTG..	23	5.12	8.50E-09	3.4
	..CAC..GAC	GTC..GTG..	17	2.61	3.60E-09	3.8
	...ACGTGA.	.TCACGT...	21	4.75	4.60E-08	2.7
assembly	GTCACGTGAC	GTCACGTGAC				

group 2	CGCCAC....	....GTGGCG	14	2.32	2.00E-07	2.1
	.GCCACA...	...TGTGGC.	21	4.06	3.30E-09	3.8
	..CCA..GTT	AAC..TGG..	23	5.86	9.10E-08	2.4
	..CCACAG..	**..CTGTGG..**	**24**	**3.53**	**2.30E-12**	**7**
	..CCA.AGT.	.ACT.TGG..	21	4.59	2.60E-08	2.9
	...CACAGT.	.ACTGTG...	24	5.41	5.20E-09	3.6
	...CAC.GTT	AAC.GTG...	24	5.79	1.90E-08	3.1
assembly	CGCCACAGTT	AACTGTGGCG				

group 3	ACC.............TGG.	.CCA.............GGT	**15**	**2.9**	**5.10E-07**	**1.7**
	.CCA.............GGT	ACC.............TGG.	**15**	**2.9**	**5.10E-07**	**1.7**
	.CCA.............TGG.	.CCA.............TGG.	22	3.16	1.10E-06	1.3
assembly	ACCA.............TGGT	ACCA.............TGGT				

isolated	CAG...TGG	CCA...CTG	**17**	**3.12**	**4.60E-08**	**2.7**

Table 1: patterns extracted by dyad-analysis with the MET family. Legends: **obs occ**: observed occurrences; **exp occ**: expected occurrences; **proba**: binomial probability; **sig**: significance index. All patterns with significance value higher than 1 were selected. Some patterns can be grouped together on the basis of sequence similarities, and assembled into larger patterns (contigs). The first group corresponds to the sequence recognized by the protein complex Met4p/Met28p/Cbf1p. The second group describes the site bound by Met31p and its homologue Met32p. These factors are those known to regulate methionine biosynthesis in the yeast *Saccharomyces cerevisiae*. To our knowledge, the third group and the isolated patterns do not show any obvious similarity to known binding sites, and could reveal new regulatory patterns.

5 Metabolic Network Analysis

We focus now on the second question, namely the functional interpretation of clusters of co-regulated genes.

Representating Metabolic Pathways as Graphs

The set of all possible metabolic reactions can be seen as a graph, with two types of nodes (metabolites and reactions respectively). Arcs represent substrate-reaction and reaction-product relationships. A graph containing all known metabolic reactions would include of the order of 10^4 nodes and as many arcs. The connectivity is very high for some particular compounds (ATP, Adenosyl-Methionine), but besides these

"pool metabolites", the vast majority of compounds are involved in a very limited number of reactions. Reactions have between 1 and 6 substrates (2 on average) and as many products. The complexity of such a graph is huge and the number of possible pathways is virtually infinite.

However, only a very restricted number of these possible pathways are effectively followed in living organisms. For instance, the database EcoCyc, which holds the most comprehensive information about *Escherichia coli* metabolism, only contains 159 distinct pathways. *E. coli* has been, for several decades, the preferred model organism for biochemists, and even though some parts of its metabolism certainly remain to be discovered, the number of pathways is not expected to increase significantly for this organism.

Metabolic Pathway Discovery

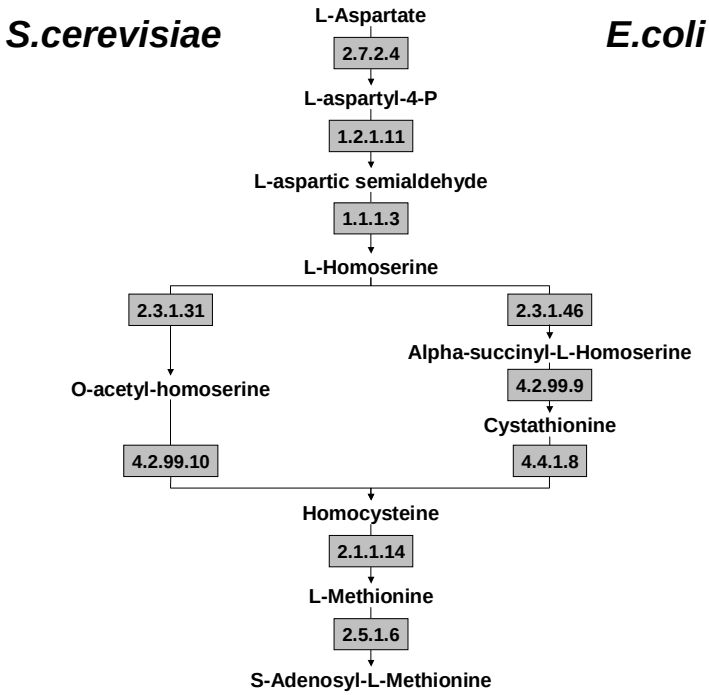

Figure 4: alternative pathways for methionine biosynthesis in *E. coli* and *S. cerevisiae*.

Many pathways however remain be discovered in other organisms. Indeed, it is common to observe that different organisms follow distinct pathways for the biosynthesis or degradation of the same molecule. For example, in *E. coli*, methionine is synthesized in 7 steps from aspartate, whereas the yeast *Saccharomyces cerevisiae* performs this transformation in 6 steps, 4 of which are common with *E. coli* (Figure

4). The case of lysine is more extreme, since *E. coli* and *S. cerevisiae* follow completely different pathways to synthesize this metabolite.

In addition, many parts of the metabolism remain largely unexplored, for example the mechanisms of toxic molecule degradation or resistance to extreme conditions observed in some bacteria.

In summary, among all the pathways that could be followed in the graph of metabolic reactions, only a very restricted fraction corresponds to already described pathways. Another part corresponds to pathways that are not yet described but might appear to be effectively used by some organisms in response to some conditions. Finally, a vast majority of these pathways might be devoid of any biological relevance. As illustrated below, measuring the transcriptional response of all the genes of an organism could be one way to select those pathways that are most likely to correspond to biological processes.

Metabolism and Gene Expression

Living organisms can rapidly modify their internal concentration of small molecules (metabolites) via enzymatic catalysis. Controlling metabolite fluxes is essential to cell viability, in that it allows the cell to maintain biochemical compounds in stationary concentrations (homeostasis), in spite of fluctuations of their external availability and internal consumption rates. Several molecular mechanisms are involved in metabolic regulation. Enzymes and transporters are regulated at different levels: transcription rate, RNA stability, translation rate, protein activity, intracellular location, protein degradation. Several of these mechanisms can be combined for the control of the same metabolic pathway. Enzymes and transporters participating in a common metabolic pathway are often co-regulated at the transcriptional level. Thus, when the culture medium is modified by depleting (or adding) a given metabolite, it is expected that the genes that participate in the biosynthesis (or degradation) of the molecule will respond at the transcriptional level. DNA chip and related technologies can be used to unravel the set of genes that respond to a given perturbation of the external conditions (addition/removal of a metabolite) in a given organism. The question is thus to discover, from this set of genes, which particular pathway could be catalyzed.

Applying Graph Analysis for a Functional Interpretation of Gene Expression Data

The first step is to select, among the set of co-regulated genes, those that code for enzymes, and identify the reactions they could catalyze. These reactions correspond to a subset of nodes in the graph of all possible metabolic reactions (Figure 5A). The method consists in trying to interconnect all these reactions in a meaningful way (Figure 5B), in order to extract a sub-graph (Figure 5C) corresponding to one or several putative metabolic pathways (Figure 5D). The algorithms for subgraph extraction and maximal path enumeration have been described elsewhere (22, 23), and we will only summarize their principle.

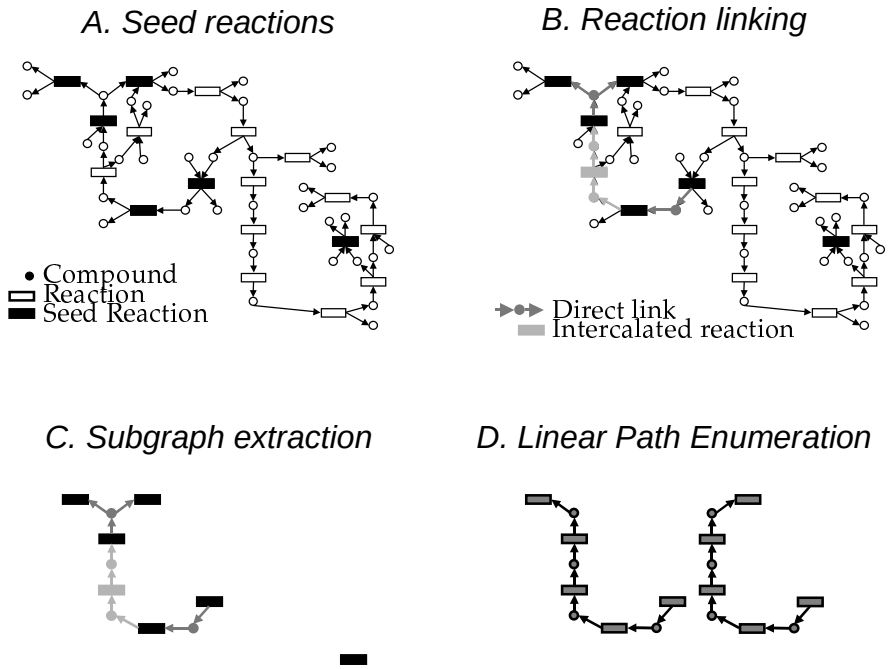

Figure 5: conceptual schema of the subgraph extraction. **A:** graph representation of metabolic pathways. Two types of nodes are used to represent metabolites (circles) and reactions (rectangles) respectively. Arcs represent the relationships between reactions and their substrates and products. Filled rectangles represent seed reactions, i.e. those that are catalyzed by genes belonging to the co-regulated cluster. **B:** A direct link (dark gray) can be established between two reactions when the first one produces a metabolite that is used as substrate by the second one. Optionally, one can decide to allow intercalating reactions (ligt gray) that were not part of the initial seeds, if this improves the connectivity. **C:** Connected components are then extracted from the graph. The extracted subgraph represents a metabolic pathway that is potentially catalyzed by the cluster of co-regulated genes. **D:** When the subgraph contains branches, it can be decomposed into non-redundant elmentary paths, which highlight potential endpoints of the metabolic pathways.

The simplest way to interconnect reactions is to identify compounds that are produced by one reaction and used as substrate by another one. In a second step, linking can be improved by intercalating reactions that were not part of the initial set. Several reasons could be invoked to justify such an intercalation. Firstly, some genes could be involved in the metabolic pathway without being regulated at the transcriptional level. Secondly, microarray technologies are still limited in reproducibility and some regulations might have escaped detection. A third possibility would be that the gene is present on the chip and its expression level has been measured correctly, but this gene has not been annotated as an enzyme yet. Indeed, for newly sequenced genomes, gene function is usually predicted by sequence similarities, and many genes remain unan-

notated. In such a case, the best candidates to ensure the missing enzymatic catalysis are the genes that belong to the initial cluster itself, but have no assigned function yet.

Comparison of Extracted Pathways with Known Pathways

Once the subgraph has been extracted, the putative pathway can be compared to the set of known metabolic pathways stored in some metabolic pathway database (9, 10, 23).

In some cases the pathway extracted from the gene cluster will correspond to some previously characterized pathway. For such cases, a simple matching of the set of reactions against a database of metabolic pathways would have provided the same answer. In other cases, one might observe only a partial match with a known pathways. The subgraph extraction might thus reveal an alternative to the pathway followed in the model organism. The method could be applied to study the metabolism of newly sequenced organisms, whose metabolism has been poorly characterized.

Finally, in some cases, one should be able to extract completely novel pathways. The co-regulation of the enzyme-coding genes would provide a good support to indicate that this pathway is biologically relevant. An interesting field of application would be to discover metabolic pathways involved in largely unexplored processes, such as resistance to toxic compounds or extreme conditions. Another application is to reveal which metabolic pathways are affected by a new drug.

Application of Pathway Analysis to the Study Case

We applied the above procedure to the 20 genes belonging to the MET cluster defined by Spellman and co-workers. Seven of these genes code for enzymes, which can catalyze 8 distinct reactions. Subgraph extraction and maximal path enumeration resulted in a linear pathway including 6 of the initial reactions (Figure 6). In this case, the linear path was obtained without intercalating any reaction that was not part of the initial set.

The pathway shows partial matches with two distinct metabolic pathways: the 4 initial steps match the sulfur assimilation pathway, and perform a progressive reduction of sulfate into sulfide. The two last steps match the methionine biosynthesis pathway, and correspond to the incorporation of sulfur into homocysteine, and the transformation of the latter into methionine. Sulfur assimilation and methionine biosynthesis are intrinsically related in the yeast Saccharomyces cerevisiae, since in this organism sulfur amino acids are all derived from the methionine biosynthesis pathway (this differs from *Escherichia coli*, where sulfur is incorporated into cysteine and then transferred to methionine). It makes thus sense to have a coordinated transcriptional regulation of all the metabolic steps from sulfate to methionine. Indeed, these genes are all known targets of the methionine-regulating transcription factors described above (18).

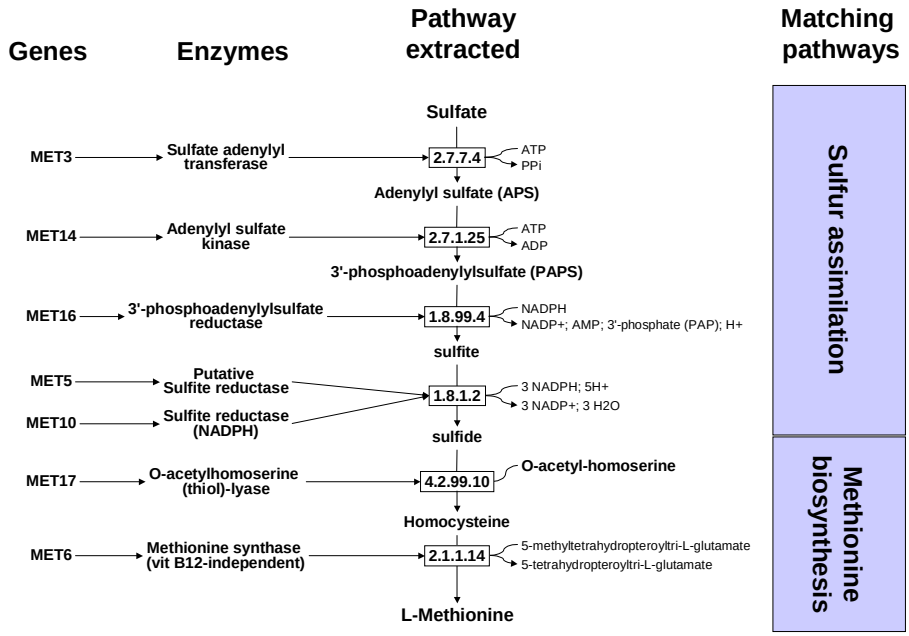

Figure 6: result obtained by pathway analysis with the cell-cycle regulated gene cluster MET from Spellman (16). Six of the reactions potentially catalyzed by enzymes-coding genes from the cluster can be assembled into a single linear pathway, without need to intercalate any additional reaction. The pathway extracted is the way used by yeast to incorporate sulfur into amino acids, by reduction of sulfate into sulfide, which is incorporated in homocyteine. This pathway matches two distinct pathways from the database: the first 4 steps correspond to sulfur assimilation, whereas the two last steps are part of the methionine biosynthetic pathway.

In summary, starting from an unordered set of reactions, the program was able to build a linear metabolic pathway, which correspond to our expectation for the study case. In this particular case, the pathway was already well characterized and a similar result would have been obtained by matching the seed reactions against a database of metabolic pathways like KEGG. However, since this pathway was re-discovered by the program without any a priori information about how reactions do assemble into pathways in the yeast (the matching with known pathways was only done a posteriori), one can hope that the same method will also provide a means of discovering novel pathway. We are currently optimizing the program and evaluating it performances in different conditions, on the basis of well characterized pathways. The optimized program will then be used to provide an interpretation of gene expression data in terms of metabolic pathways.

6 Conclusions

In the context of genomic approaches, coding sequence analysis is often insufficient to systematically assign a function to each gene. The function depends not only on the structure of the encoded protein, but also on the context in which this protein exerts its activity. Functional predictions thus require the integration of different levels of information.

The possibility to measure the transcriptional response at a genome scale offers exciting perspectives for the discovery of gene function, taking into account the ways genes are associated in functional clusters. By combining regulatory sequence analysis and metabolic pathway analysis, one could obtain two independent and complementary sources of information for these clusters of co-regulated genes. The same methods also apply to clusters of genes obtained from other functional genomics approaches, such as phylogenetic profiles (13) and gene fusion/fission analysis (6, 11, 12).

7 Availability

Regulatory Sequence Analysis tools are available on the web (20) at the URL http://ucmb.ulb.ac.be/bioinformatics/rsa-tools/. The home page for the EBI project of database on Protein Function and Biochemical Pathways is at http://www.ebi.ac.uk/research/pfbp/. A prototype version of the patwhay analysis tools can be accessed from this site. A prototype version of the visualization tools is available at http://www.soi.city.ac.uk/~msch/.

8 Acknowledgements

Jacques van Helden was funded by the European Commission Contract N0: QLRI-CT-1999-01333.

References

1. Brazma, A., I. Jonassen, J. Vilo, and E. Ukkonen. 1998. Predicting gene regulatory elements in silico on a genomic scale. Genome Res 8:1202-15.
2. Brown, P. O., and D. Botstein. 1999. Exploring the new world of the genome with DNA microarrays. Nat Genet 21:33-7.
3. DeRisi, J. L., V. R. Iyer, and P. O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278:680-6.
4. Eisen, M. B., and P. O. Brown. 1999. DNA arrays for analysis of gene expression. Methods Enzymol 303:179-205.

5. Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95:14863-8.
6. Enright, A. J., I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. 1999. Protein interaction maps for complete genomes based on gene fusion events [see comments]. Nature 402:86-90.
7. Gilbert, D., M. Schroeder, and J. van Helden. 2000. Interactive visualization and exploration of relationships between biological objects. Trends in Biotechnology accepted.
8. Graber, J. H., C. R. Cantor, S. C. Mohr, and T. F. Smith. 1999. Genomic detection of new yeast pre-mRNA 3'-end-processing signals. Nucleic Acids Res 27:888-94.
9. Kanehisa, M., and S. Goto. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28:27-30.
10. Karp, P. D., M. Riley, M. Saier, I. T. Paulsen, S. M. Paley, and A. Pellegrini-Toole. 2000. The EcoCyc and MetaCyc databases. Nucleic Acids Res 28:56-59.
11. Marcotte, E. M., M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science 285:751-3.
12. Marcotte, E. M., M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg. 1999. A combined algorithm for genome-wide prediction of protein function [see comments]. Nature 402:83-6.
13. Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96:4285-8.
14. Reinert, G., and S. Schbath. 1998. Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. J Comput Biol 5:223-53.
15. Salgado, H., A. Santos-Zavaleta, S. Gama-Castro, Z. r. D. Mill#n, F. R. Blattner, and J. Collado-Vides. 2000. RegulonDB (version 3.0): transcriptional regulation and operon organization in Escherichia coli K-12. Nucleic Acids Res 28:65-67.
16. Spellman, P. T., G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9:3273-97.
17. Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A 96:2907-12.
18. Thomas, D., and Y. Surdin-Kerjan. 1997. Metabolism of sulfur amino acids in Saccharomyces cerevisiae. Microbiol Mol Biol Rev 61:503-32.
19. van Helden, J., B. Andre, and J. Collado-Vides. 1998. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol 281:827-42.
20. van Helden, J., B. Andre, and J. Collado-Vides. 2000. A web site for the computational analysis of yeast regulatory sequences. Yeast 16:177-87.
21. van Helden, J., M. del Olmo, and J. E. Perez-Ortin. 2000. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. Nucleic Acids Res 28:1000-10.

22. van Helden, J., D. R. Gilbert, L. Wernisch, and S. Wodak. 1999. Logical tools for querying and assisting annotation of a metabolic and regulatory pathway database. ISMB .
23. van Helden, J., A. Naim, R. Mancuso, M. Eldridge, L. Wernisch, D. Gilbert, and S. Wodak. 2000. Representing and analysing molecular and cellular function in the computer. accepted in Biological Chemistry .
24. van Helden, J., A. F. Rios, and J. Collado-Vides. 2000. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res 28:1808-18.
25. Vilo, J., A. Brazma, I. Jonassen, and E. Ukkonen. 2000. Mining for Putative Regulatory Elements in the Yeast Genome Using Gene Expression Data. ISMB .
26. Wingender, E., X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Pr, I. Reuter, and F. Schacherer. 2000. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res 28:316-319.