Network inference from gene expression profiles: what "physical" network are we seeing? (Prokaryotes vs Eukaryotes)

M. Zampieri, N. Soranzo, D. Bianchini and C. Altafini
SISSA-ISAS, International School for Advanced Studies
via Beirut 2-4, 34014 Trieste, Italy

November 14, 2007

Abstract

The concept of reverse engineering a gene network, i.e., of inferring a genome-wide graph of putative gene-gene interactions from high throughput microarray data has been used extensively in the last years to deduce/integrate/validate various types of "physical" networks of interactions among genes or gene products. This paper investigates which of these networks emerge significantly when reverse engineering collections of gene expression data for two model organisms, *E.coli* and *S.cerevisiae*, without any prior information. For the first organism the pattern of expression correlations is shown to reflect in fine detail both the operonal structure of the DNA and post-transcriptional regulatory effects on the gene products, *in primis* the co-participation in a protein complex, while for the second organism we find that direct transcriptional control (e.g., transcription factor – binding site interactions) has no statistical significance and also post-transcriptional regulatory mechanisms (such as co-sharing a protein complex, colocalization on a metabolic pathway or compartment) are resolved at a lower level of detail, thereby reflecting the different complexities of the two organisms.

Manuscript length: approximately 48000 characters (including figures)

Reverse engineering a gene network means extrapolating a graph of putative gene-gene interactions from high throughput microarray data. Many are the algorithms that have been proposed for this scope in recent years (see [2, 6, 9] for an overview) and many the (very) different contexts of application: deduce/integrate/validate various types of "physical" networks of interactions between genes or gene products, see e.g. [1, 3, 7, 11, 12, 13, 15, 17, 20, 25, 29].

Our aim in this paper is to address the following (related) question: what among these different networks is more likely to emerge from a completely unsupervised reverse engineering processing of the gene expression data and at which level of detail can we confidently reconstruct such networks? In other words: what is the more likely biological origin of the pattern of gene-gene expression similarities we see probing only the "layer" of transcripts without adding any a priori information neither on the "upstream" regulatory interactions (like a direct transcriptional activation could be considered) nor in the "downstream" one (at the level of protein or of metabolic interactions)?

For this purpose, we collected a number of possible alternative networks for the two model organisms: maps of transcription factors – biding sites (TF-BS), protein – protein interactions (PPI), protein complexes (PC), and metabolic pathways (MP). In order to take into account also the architecture of the genomes, we considered maps of paralog genes (PAR) [28] and, for *E.coli* alone, a map of transcription units (TU) describing the operonal structure of the prokaryotic DNA (see Tables (a) and (b) of Fig. 1 and Supplementary Notes for details and data sources). As for gene profiling, we used three different datasets: one for *E.coli* (445 Affymetrix experiments) and two for *S.cerevisiae* (one containing 958 cDNA experiments, the other 790 Affymetrix experiments). For this last organism, as a byproduct, the comparison of the two datasets allows to evaluate the differences between the two gene profiling technologies (see in particular Fig. 1 and Fig. 3).

Examples of how to conjugate gene expression with one of these physical networks are [7] and [17] where expression similarity (together with sequence compatibility) is used to infer new putative TF-BS edges. Rather than TF-BS, the same comparison between expression similarity and a given network graph can alternatively lead to putative new PPI edges [14, 15, 23]. As a matter of fact, according to [25], for *S.cerevisiae*, gene expression correlation is the most significant among the 17 indexes considered for this scope (including, among others, ontological information, sequence similarity, protein localization and domain structure, etc.). Similar uses of gene expression have been published in the context of metabolic pathways: see e.g. [13, 21], or to predict prokaryotic operonal structure [11, 26]. Needless to say, the integration of several of the "physical" maps above is one of the very often used approaches in the literature [10, 18, 20, 27, 32].

There are several motivations that justify the simultaneous use of gene expression in these and other biological contexts, the first and foremost being that genes, gene products and metabolites form a unique complex interlinked system, whose unraveling is far from complete, especially for what concerns its context-dependence (condition-specific activation of regulatory mechanisms, dynamic behavior, dependences from internal and external parameters such as nutrients and stimuli, etc.).

Another reason is that the gene expression "layer" is the only one that can be measured in such a systematic way. A third reason is that even zooming to this layer alone, the current amount, quality and significance of microarray data is drastically insufficient.

Overrepresented networks comparison

Assuming no prior knowledge, a network structure can be inferred solely from microarray data by means of a "similarity matrix" [4] (see Supplementary Notes for definitions and algorithms) and used to test which of the types of interactions listed in Fig. 1 is significantly visible. For the two organisms, the edges weights resulting from the statistical analysis are rank-ordered and the percentages of "true" edges of each physical network in the top 1% of edges are shown in the histograms of Fig. 1. For *E.coli* (Fig. 1(c)) we observe that more than 50% of the TU map is detected (against a random pick of 1%) meaning that the pattern of expression similarity is strongly influenced by the operonal structure of the DNA, as is well-known [11, 26]. The other emerging network, the (manually curated) protein complexes, is relevant also for *S.cerevisiae* (Fig. 1(d)). Notice how in *S.cerevisiae* the percentage decreases drastically passing from the manually curated protein complexes (PC1) to the complexes identified by means of systematic screening (PC2). This consideration extends to PPI on both organisms: the protein-protein bindings detected by high throughput essays need not correspond to stable bindings and hence to highly correlated patterns of expression. On both organisms the direct transcriptional regulation due to the transcription factors (TF-BS map) is far from being the most relevant indicator. However, while for *E.coli* it remains in the range of significance of other networks (around 6-8%, like MP), in *S.cerevisiae* the map TF-BS is below the threshold of statistical relevance (for a q.value of 0.05, see Supplementary Notes) in both datasets we collected. Concomitant causes such as combinatorial regulatory effects [1] or condition-specific activation of the TF-BS edges [17, 24] certainly play a role in the loss of relevance of this class of interactions.

Clustering the data

The edges of highest significance, suitably clustered, can be tested against the most relevant physical networks emerging from the previous analysis. In particular, the correspondences clusters/PC for the two organisms are shown in Fig. 2.

E.coli For *E.coli*, the clustered expression correlations reproduce faithfully a large part of the collection of PC, and the matching clusters-PC is quasi-monogamous (see also Fig. S4 and Fig. S5 for details and statistics). A similar (even better) univoque correspondence is detected between the clusters and the TU (see Fig. S6, and the statistics in Fig. S7), while for MP the percentages

are lower but still significant (see Fig. S8 and Fig. S9). Most often co-clustered genes share similar functional annotation (see Fig. 3) and can be used to infer/confirm biological hypothesis.

A thorough description of the ontological information deduced from the clusterization procedure is provided in the Supplementary Notes. The most striking example is represented by the largest cluster, which includes (in 61 genes) basically all the 50 genes known to be involved in flagellar formation and function. Apart from the flagellum complex subunits (24) and its transcriptional regulators (flhDC and the σ^{28} factor fliA), the cluster contains chemotactic genes, genes regulated by the flhDC complex, by the σ^{28} factor or the anti-σ^{28} factor, other genes involved in flagellar biogenesis and motility, or predicted regulators of the σ^{28} factor. Such a functional compactness (and disconnection from the rest of the gene network, see Fig. S3) probably originates from *E.coli*'s need to activate the flagellum in every kind of experimental condition and in constant stoichiometric ratio. Also ribosomal genes tend to form large clusters of functionally similar genes (mainly concentrated in clusters 10, 20 and 25) going beyond the operonal structure and forming different ribosomal structural components (rpl, rps, rpm, rpo). Another remarkably homogeneous set of genes not induced by any operon is in cluster 24: of its 10 genes, 9 are associated with the SOS pathway.

The list of significant clusters is long, as essentially all basic functions needed for survivial and growth are captured by the clusterization. Nucleotide (cluster 56 for pyrimidine, cl. 88 for purine) and aminoacid biosynthesis are recurrent biological functions retrieved by the procedure. For this last function, the resolution is often at the level of the single aminoacid, like serine biosynthesis and threonine biosynthesis from homoserine (cl. 7), tryptophan and histidine biosynthesis (cl. 5), arginine biosynthesis (cl. 36), methionine biosynthesis (cl. 69, 7), alanine biosynthesis (cl. 404), isoleucine biosynthesis from threonine (cl. 72) and cysteine biosynthesis (cl. 9). The single resolution extends to tRNAs: valine tRNAs (cl. 171), glutamate tRNA (cl. 175), asparagine tRNA (cl. 102), methionine tRNA (cl. 166), glycine tRNA (cl. 167), leucine tRNA (cl. 168), although sometimes similar enzymatic functions prevail (like in cluster 41 where genes involved in aminoacid-tRNA synthetase for five different aminoacids are grouped).

Biosynthetic pathways are visible for many (other) compounds, like, for example, thiamine (cl. 21), enterobactine (cl. 14), spermidine (cl. 133), etc. Likewise for degradatory pathways (e.g. alanine in cl. 404, threonine in cl. 185, L-arabidose in cl. 26, etc.), and for many elements of the superfamily of ABC transporters.

Well detected are the responses to various stresses, like osmotic (cl. 80, 139), oxidative (cl. 415), thermal (cl. 106, 184), acid (cl. 308) and extracytoplasmatic (cl. 340). Also metabolic functions, like for example aerobic and anaerobic respiration, are well identified by specific and disjoint clusters. For instance for the aerobic respiration, cluster 34 contains the sdhCDAB-sucABCD operon involved in the two consecutive succinate-related steps of the TCA Cycle. A cluster related to anaerobic respiration is cluster 117, which contains part of the fixABCX TU, thought to be

involved in the anaerobic metabolism of carnitine. This last hypothesis is reinforced by the co-clusterization with caiD, a gene having a carnitine racemase activity. Significant is also cluster 203, containing 3 genes belonging to three different TU but all involved in the anaerobic respiration. The preferred electron acceptor for anaerobic respiration in *E.coli* is nitrate that is reduced to nitrite which is either excreted or further reduced. *E.coli* contains 3 nitrate reductases: two of them, nitrate reductase A (NRA) and nitrate reductase Z (NRZ), are membrane bound, while the third one, Nap, is located in the periplasm. Their different environmental conditions for activation are reflected in the formation of three separate and neatly defined clusters (cl. 98, 233, 140). Similar considerations extend to the 2 nitrite reductases (cl. 57 and 246). In addition, nitrate serves as a nitrogen source, an important constituent of protein and amino acids, and nitrogen metabolism is a function that emerges compactly from our analysis (cl. 3). Iron transport is usually involved in the formation of proteins belonging to the respiration chain, as it has an electron acceptor activity and is represented here by clusters 19. Assimilation of other substrates such as sulfur and carbon are depicted respectively by clusters 9, 19, 347, and 46, 291, 393.

Several other clusters contain clues about putative gene functions, like cluster 67 encoding for two components of the dmsABC, Dimethyl sulfoxide (DMSO) reductase, a terminal electron transfer enzyme functioning anaerobically in absence of nitrate. The other genes in the cluster are paralogs, like, ynfF and ynfE (highly similar to dmsA), ynfG (highly similar to dmsB), and ydfZ. Little is known about ydfZ, but the working hypothesis [16] is that it is activated under anaerobic growth, and the clusterization reinforces this assumption. Another example of biological inference is cluster 161. It contains sgcABC, part of the sugar transporting phosphotransferase system (PTS), together with ytfT, that, although part of a different TU, according to sequence similarity may function as an ATP-dependent sugar transporter, hypothesis consistent with our results.

The operonal structure of the genome is certainly a key factor in the formation of the clusters, but alone does not exhaust the information that can be extrapolated from the expression correlation patterns, see Fig. 2 and Fig. 4. We can notice for instance that the distribution of intracluster average gene distances (shown in Fig. 4(b)) although largely comparable to that of the TU, has a heavier tail, more related to the PC distribution. Most of the large clusters are examples of functional information not exhausted by any operonal structure. It is interesting to notice that the difference in the overlap clusters/TU concerns most often the genes located at the boundaries of the operons (see e.g. cl. 3, 5, 6, 10, and many more). As a confirm that the operonal structure and/or protein complex interactions are much stronger mediators of co-expression than direct DNA binding (i.e. being a pair of TF-BS), we notice that co-clusterization of these last pairs are sporadic (e.g. cl. 1, 3, 7, 24, 38, 74, 101).

S.cerevisiae The clusterization procedure is repeated also for *S.cerevisiae* (see Supplementary Notes for details). As can be seen in Fig. 2, while the correspondence clusters-complexes (of type PC1) is still acceptable, the percentages of subunits detected for the complexes are drastically

reduced with respect to *E.coli*. Also qualitatively, the inferred results are quite different, with a few very accurate reconstructions of large complexes but much less information content in the medium-small size clusters. Large and small ribosomal subunits are captured very precisely for both cytoplasmic (cl. 1) and mitochondrial (cl. 3) ribosomes. This last cluster (of 70 genes) is a good example of compartmental homogeneity: the 56 mitochondrial ribosomal genes are in fact co-clustered with 6 more genes from the mitochondrial membrane translocases. Even more compact clusters (in terms of both localization and function) are cluster 6, with 25 of the 32 subunits of the proteasome (in 34 genes of the cluster), and cluster 5, which contains all the respiratory chain complexes (34 in 36 genes of the cluster). Notice how in this last case also the main transcriptional regulator of the oxidative phosphorylation (HAP4) is co-clustered, one of the very few examples of TF-BS edges detected. In general, the large clusters tend to co-localize but also to share complex subunits (see the example of the RNA polymerases complexes scattered in clusters 2, 4, and 7). As for the remaining medium-small size clusters, most of those having a significant annotation tend to be involved in transcription and translation processes, while metabolic functions are fragmentary and do not emerge from the clusterization, mostly because many enzymatic genes are missing (they have no significant correlation coefficients). For example two pairs of enzymes of Glycolysis are co-clustered in cluster 8, but most of the other genes in the pathway are not passing the correlation filter. A few clusters containing eminently metabolic genes are however present (e.g. cl. 12, 15, 21, 30, 31, 100), although they are not pathway-specific. Sometimes genes co-localize also in other compartments like the endoplasmatic reticulum (15), the cytoskeleton (37) or the golgi vescicle (117).

An example of how to use the clusterization in the verification of uncertain functional annotations is the following. The gene PPE1 (YHR075C, also known as MRPS2) among other annotations, is also identified as a small subunit mitochondrial ribosomal protein [31, 8], an annotation which is contradictory with e.g. the results of [30]. In our analysis PPE1 is lost at the correlation filter, meaning that it has no strong and stable interaction with any other gene. Extending for example to the 10 "newly" reported subunits of mitochondrial ribosomes of [8], 7 are correctly included in cluster 3 and 1 in cluster 8 (still mitochondrial) and only 2 are missing (YMR158W and YPL013C).

Conclusion

The systematic observation of the patterns of gene coexpression tends to unveil functional categories that are stable rather than transient or condition-specific [29]. For them, the picture emerging from the genome-wide analysis in the two organisms shows common aspects, like the coexistence of various "layers" of regulation, or the importance of post-transcriptional interactions among the gene products (coparticipation in a complex, colocalization, etc.), but also a marked decrease into the visibility of the direct transcriptional control when passing from the prokaryotic to the eukaryotic genome. The increase in complexity of regulatory mechanisms, genome architecture and number

of functions per gene is inversely proportional to the ability of retrieving significant and detailed information by means of a reverse engineering approach.

Materials and methods (including statistical analysis)

see Supplementary Notes.

References

[1] S. Balaji, M.M.Babu, L. Iyer, N. Luscombe, and L. Aravind. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol.*, 360:213–27, 2006.

[2] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo. How to infer gene networks from expression profiles. *Mol Syst Biol*, 3(78), 2007.

[3] K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. Reverse engineering of regulatory networks in human B cells. *Nat. Genet.*, 37:382–390, 2005.

[4] A. J. Butte and I. S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, pages 418–429, 2000.

[5] Y. Chen, C. Chou, X. Lu, E. Slate, K. Peck, W. Wu, and J. Voit, E.O. Almeida. A multivariate prediction model for microarray cross-hybridization. *BMC Bioinformatics*, 7:101, 2006.

[6] H. De Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9:67–103, 2002.

[7] J. Faith, B. Hayete, J. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. Collins, and T. Gardner. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5:54–66, 2007.

[8] X. Gan, M. Kitakawa, K. Yoshino, N. Oshiro, K. Yonezawa, and K. Isono. Tag-mediated isolation of yeast mitochondrial ribosome and mass spectrometric identification of its new components. *Eur. J. Biochem.*, 269:5203–5214, 2002.

[9] T. S. Gardner and J. J. Faith. Reverse-engineering transcriptional control networks. *Physics of Life Rev.*, 2:65–88, 2005.

[10] M. J. Herrgård, M. W. Covert, and B. Palsson. Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res.*, 13:2423–2434, 2003.

[11] R. Hershberga, E. Yeger-Lotema, and H. Margalit. Chromosomal organization is shaped by the transcription regulatory network. *Trends in Genetics*, 21(3):138–142, 2005.

[12] D. Hwang, A. G. Rust, S. Ramsey, J. J. Smith, D. M. Leslie, A. D. Weston, P. de Atauri, J. D. Aitchison, L. Hood, A. F. Siegel, and H. Bolouri. A data integration methodology for systems biology. *Proc Natl Acad Sci U S A*, 102(48):17296–17301, 2005.

[13] J. Ihmels, R. Levy, and N. Barkai. Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nature Biotech.*, 22(1):86–92, 2004.

[14] R. Jansen, D. Greenbaum, and M. Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome Res*, 12(1):37–46, 2002.

[15] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science*, 302:449–453, 2003.

[16] Y. Kang, K. Weber, Y. Qiu, P. Kiley, and F. Blattner. Genome-wide expression analysis indicates that FNR of Escherichia coli K-12 regulates a large number of genes of unknown function. *J Bacteriol.*, 187:1135–60, 2005.

[17] H. Kim, W. Hu, and Y. Kluger. Unraveling condition specific gene transcriptional regulatory networks in Saccharomyces cerevisiae. *BMC Bioinformatics*, 7:165, 2006.

[18] J. Korbel, L. Jensen, C. von Mering, and P. Bork. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat. Biotechnol.*, 22:911–917, 2004.

[19] R. Kothapalli, S. Yoder, S. Mane, and T. Loughran. Microarray results: how accurate are they? *BMC Bioinformatics*, 3:22, 2002.

[20] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte. A Probabilistic Functional Network of Yeast Genes. *Science*, 306(5701):1555–1558, 2004.

[21] Z. Li and C. Chan. Integrating gene expression and metabolic profiles. *J Biol Chem*, 279(26):27124–27137, Jun 2004.

[22] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15-th International Conference on Machine Learning*, pages 296–304, 1998.

[23] L. J. Lu, Y. Xia, A. Paccanaro, H. Yu, and M. Gerstein. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res*, 15(7):945–953, Jul 2005.

[24] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431:308–312, 2004.

[25] Y. Qi, Z. Bar-Joseph, and J.Klein-Seetharaman. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, 63:490–500, 2006.

[26] C. Sabatti, L. Rohlin, M.-K. Oh, and J. C. Liao. Co-expression pattern from dna microarray experiments as a tool for operon prediction. *Nucleic Acids Research*, 30:2886–2893, 2002.

[27] N. Simonis, J. van Helden, G. N. Cohen, and S. J. Wodak. Transcriptional regulation of protein complexes in yeast. *Genome Biol*, 5(5), 2004.

[28] S. Teichmann and M. Babu. Gene regulatory network growth by duplication. *Nat. Genet.*, 36:492–496, 2004.

[29] S. A. Teichmann and M. M. Babu. Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends in Biotechnology*, 20(10):407–410, Oct 2002.

[30] B. P. Tu, A. Kudlicki, M. Rowicka, and S. L. McKnight. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science*, 310(5751):1152–1158, 2005.

[31] J. Wu, T. Tolstykh, J. Lee, K. Boyd, J. B. Stock, and J. R. Broach. Carboxyl methylation of the phosphoprotein phosphatase 2a catalytic subunit promotes its functional association with regulatory subunits in vivo. *EMBO J*, 19(21):5672–5681, 2000.

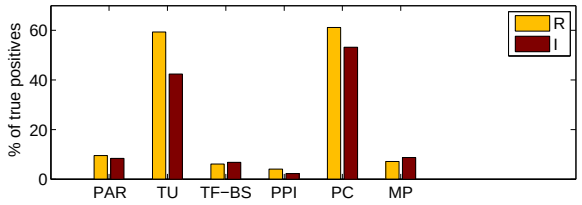
[32] Y. Yamanishi, J. P. Vert, and M. Kanehisa. Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, 20 Suppl 1:363–370, Aug 2004.

physical interaction network	acronym	n., type of edges
paralog genes, SW> 1000	PAR	714, undirected
transcription units	TU	7052, undirected
transcr. factors - binding sites	TF-BS	3071, directed
protein - protein interactions	PPI	33324, undirected
protein complexes	PC	2228, undirected
metabolic pathways	MP	3804, directed

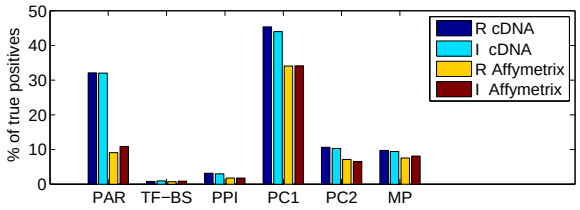
(a) Various "physical" networks collected: *E.coli*

physical interaction network	acronym	n., type of edges
paralog genes, SW> 1000	PAR	4268, undirected
transcr. factors -binding sites	TF-BS	12376, directed
protein - protein interactions	PPI	23278, undirected
protein complexes, annotated	PC1	21616, undirected
protein complexes, systematic	PC2	120110, undirected
metabolic pathways	MP	4471, directed

(b) Various "physical" networks collected: *S.cerevisiae*

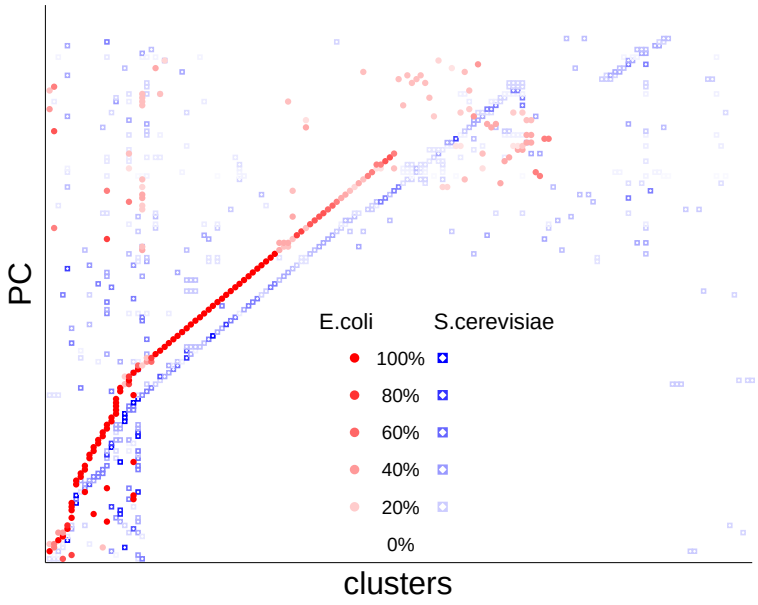


(c) Networks represented in the top 1% of inferred edges for *E.coli*

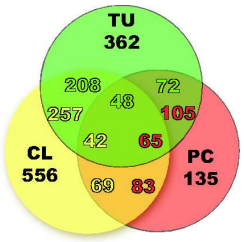


(d) Networks represented in the top 1% of inferred edges for *S.cerevisiae*, for both cDNA and Affymetrix data

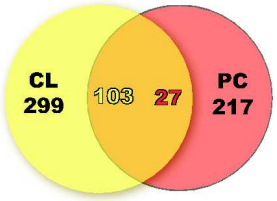
Figure 1: **Overrepresented physical networks.** For each of the two organisms we collected several networks representing different genomic or physical interaction properties, shown in Table (a) and (b), see Supplementary Notes for data sources. The similarity matrices, computed with Pearson correlation (R) and with mutual information (I) and representing the predicted likelihood of an edge between any two genes, are compared with the graphs of the various networks listed in Tables (a) and (b). Since the ratio between number of experiments and number of genes is very low (around 0.1), the inference power is also very low. The "coarse grain" statistics we use to describe the results are obtained sorting the inferred weights, binning them into 100 bins and counting the percentage of "true" edges (of each physical network) lying in each bin. The percentages of true positives in the top bin are shown in the histograms (a randomly chosen network would yield 1% of true positives). In absolute terms, the degree of inference remains very low, as each bin contains a huge number of edges (94373 for *E.coli* and 192355 for *S.cerevisiae*). However, if we are interested only in comparing the fitting between the various physical networks, the differences in overrepresentation in the highest weight bin are a reasonably objective metric. (c): *E.coli* inference. Two networks are neatly emerging, TU and PC. The first emphasizes the visibility in the expression pattern of the operonal structure of the DNA. The TU and PC detected have an overlap which is consistent but still below 50% (of the 2632 TU edges and 1364 PC edges in the top 1%, 694 are in common), meaning that also co-participation in a PC is a strong, independent source of coexpression. (d): *S.cerevisiae* inference, cDNA and Affymetrix data. The dominant index is PC1 in both datasets, followed by the map of duplicated genes. The high magnitude of the two peaks in the cDNA data alone strongly suggests that this technology may be affected by a systematic bias towards aspecific binding and cross-hybridization of genes with sequence similarities [19, 5], see also Fig. 3. With the exception of TF-BS for *S.cerevisiae*, all histograms are statistically significant (q.value < 0.05, see Supplementary Notes and Fig. S2).



(a) Clusterization vs PC

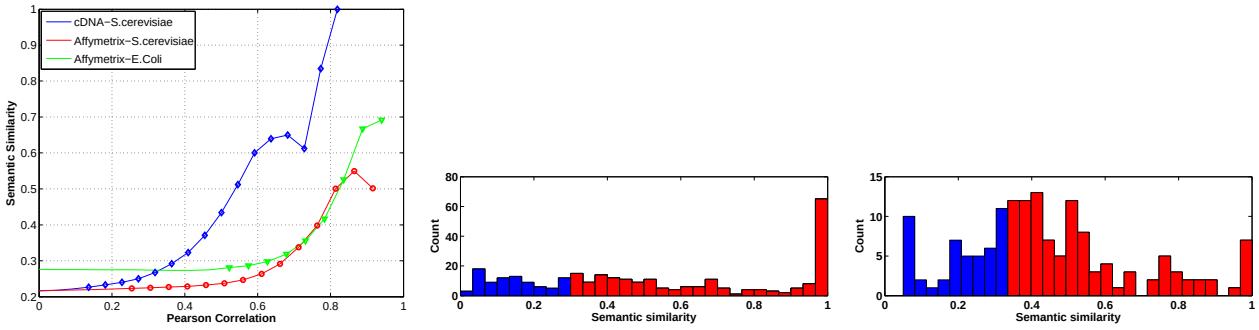


(b) *E.coli*



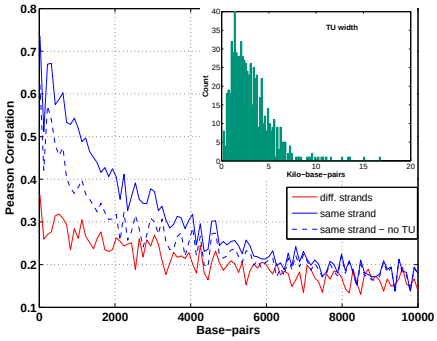
(c) *S.cerevisiae*

Figure 2: **Correspondences between expression clusters and protein complexes for *E.coli* and *S.cerevisiae*.** The edges of highest correlation, selected balancing graph coverage and connectivity degreee and suitably clustered (see Supplementary Notes for details), are checked against the protein complexes (PC1 for *S.cerevisiae*). In (a) the correspondences clusters/PC are shown, with the brightness scale indicating the percentage of genes of each PC in the cluster. In *E.coli*, PC are consistently and "monogamously" matched (see Fig. S4 for a more detailed representation and Fig. S5 for a statistical analysis). In *S.cerevisiae*, while the clusterization is still sufficiently accurate, the most significant difference is in the percentage of complex subunits detected in average by the thresholding, implying that the complexes have a lower degree of cohesion in terms of gene expression. Full detail on the complexes and a few statistical parameters are provided in Fig. S11 and Fig. S12. The Venn diagram for *E.coli* shows how many groups of genes of one of the three categories, clusters, TU and PC, are completely contained in the groups of the other two (monocromatic inclusion: a group of genes of type X belongs to a single group of type Y, see Fig. 1 for the TU/PC overlap with a more relaxed criterion). For example there are 72 TU contained in the 135 PC, and 105 PC contained in the TU. Of these 105, 65 are completely included simultaneously in TU and clusters. For what concerns the ability of the clusterization to infer PC and TU, if in absolute terms the correspondence clusters/TU is certainly higher, in percentage it is of the same order (61% for PC and 57% for TU). These percentages are much higher than in *S.cerevisiae* (10%).

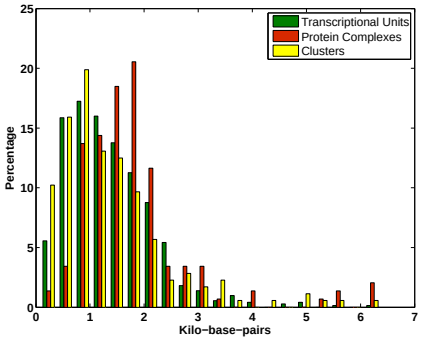


(a) (b) (c)

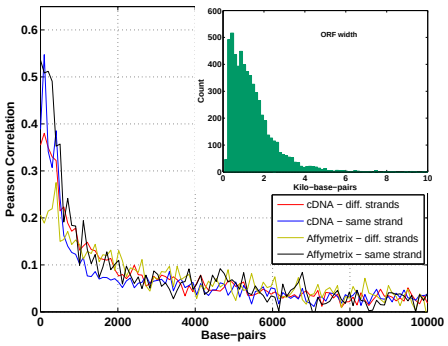
Figure 3: **Pearson correlation and semantic similarity**. We used a quantitative measure of semantic similarity between gene products [22] (see Supplemetary Notes) in order to evaluate whether genes with similar function share similar expression profiles. When comparing semantic similarity with coexpression, (a), we see that rather than organism-specific, the differences are platform-specific. If for Affymetrix data the two graphs are similar, the curve grows much faster for cDNA data. This seems to be due to the more aspecific hybridization that characterizes cDNA chips: since genes are often annotated according to sequence similarity, the cross-hybridization bias is amplified towards highly co-regulated pairs [19]. The peak in corrispondence of the maximal intracluster semantic similarity in *E.coli*, (b), reflects the matching clusters/operons and is missing in *S.cerevisiae*, where however a sufficiently high degree of functional homogeneity still characterizes the majority of the clusters (bins in red have p.value $\leqslant 0.05$, see Supplementary Notes).



(a) *E.coli*: coexpression v.s. distance (inset: TU width)



(b) *E.coli*: intracluster average distance



(c) *S.cerevisiae*: coexpression v.s. distance, cDNA and Affymetrix datasets (inset: ORF width)

Figure 4: **Pearson correlation and distance on the genome.** Coexpression decays more rapidly with distance in *S.cerevisiae* than in *E.coli*: the correlation drops to 0.2 at a distance of 6 Kbp in E.coli (a), as opposed to 1 Kbp in *S.cerevisiae*, for both cDNA and Affymetrix datasets (c). In *E.coli* the value 6 Kbp is consistent with the distribution of TU width (inset panel in (a)). Genes on the same strand have much higher correlation than genes on opposite strands. For *E.coli*, even if we restrict to gene pairs not involved in a TU (see dashed blu line in (a)), the influence of distance on coexpression is still clearly visible. In *S.cerevisiae*, the short-range high correlation peak is represented almost completely by overlapping ORFs (the distribution of ORF widths is shown in the inset), for which the cDNA experiments cannot discern any strand-specificity, unlike Affymetrix experiments. In panel (b), the distribution of intracluster average distances (see Supplementary Notes) for *E.coli* is compared with the corresponding distributions of average distances among PC and TU subunits. The histogram for the clusters is more similar to that of TU than PC, although its tail is heavier are more related to PC. A similar analysis is impossible for *S.cerevisiae* as the vast majority of clusters is composed of genes located on different chromosomes.