A network characterization of metabolic flux predictions, medium-dependent essentiality and metabolic inconsistency

by

Nikolaus Sonnenschein

A thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
in *Bioinformatics*

Approved, Thesis Committee:

Prof. Dr. Marc-Thorsten Hütt
Jacobs University Bremen

Prof. Dr. Georgi Muskhelishvili
Jacobs University Bremen

Dr. Arndt Benecke
Institut des Hautes Études Scientifiques

Date of Defense, January 7, 2011

School of Engineering and Science

Für Thilo

Abstract

The thesis presented here is summarizing and interconnecting a series of articles on the network organization of metabolic and gene regulatory processes. These publications are the product of my PhD work. They cover the following scientific topics:

(i) Does the spatial distribution of genes on a chromosome deviate from randomness, and if so, does it contribute to the transcriptional regulation of genes? Using methods from point-process statistics, we have been able to show that genes regulated by transcription factors segregate from genes with no such regulatory information available, which constitute the vast majority of genes on the chromosome of *Escherichia coli*. This evolutionary "unmixing" of the two classes of genes is evidence of the involvement of chromosome organization in processes of gene regulation and draws a connection to the following topic (ii).

(ii) Is the transcriptional regulation of metabolic processes predominantly accomplished by the classically conceived form of control that is exerted by the *digital* actions of transcription factors, or does the homeostatic control of chromosome and chromatin structure play a significant role, which is an analog type of regulation? In *Escherichia coli* we find a strong coherence of gene expression changes and metabolic network topology. This coherence is destroyed upon perturbation of the *analog* part of the regulatory machinery. We argue that this is evidence for an *analog* regulation of metabolic demand.

(iii) Is it generally possible to quantify the impact of perturbations to the regulatory machinery and environment of an organism by the integration of gene expression profiles and genome-scale metabolic reconstructions? For gene expression profiles obtained from adrenocortical adenomas, we compare the coherence developed for the investigation in topic (ii) with an inconsistency measure developed by other groups, which employs constraint-based analysis. We find a strong negative correlation between both approaches, indicating that a large part of the more sophisticated constraint-based measure can be interpreted topologically, i.e, from a network perspective. In a detailed view on the inconsistency we are able to extract valuable, physiological information from the otherwise difficult to access expression data.

(iv) Can we assign topological markers to different forms of perturbations in the systems under investigation, e.g., the medium-dependent essentiality of reactions or metabolically altered system states in cancer cells? Using methods from graph theory and constraint-based modeling we study reaction essentiality as well as the inconsistency in topic (iii) from a network perspective.

An introduction to these topics is provided in the first part of this thesis. In addition, concluding remarks and a future outlook will be given at the end of the thesis.

Contents

Chapter 1

Introduction

Metabolism and gene regulation

Metabolism constitutes the system of chemical reactions concurrently taking place inside a living cell. The cell acts hereby as an open system [129], constantly exchanging matter with its surroundings by taking up building blocks and energy yielding substances as well as getting rid of waste material. The fact that it is an open system permits, under certain circumstances, the occurrence of steady states, i.e, time-independent dynamic equilibria (*Fließgleichgewichte*) under which the composition of the system stays invariant despite the continuous exchange of its components, in that sense, being fundamentally different to the classical notion of a thermodynamic equilibrium. It is by favoring the right circumstances, that evolution has brought about life.

As environments constantly fluctuate, robust control mechanisms are necessary for metabolism to maintain homeostasis, i.e., to quickly regain a steady state upon perturbation. Because most reactions are catalyzed by enzymes, one major form of regulation is exhibited by controlling their concentrations and activities. Gene regulation is thus one of the major components in the robust control of metabolism. Metabolism and gene regulation are the key issues of this thesis, and we will approach both topics from multiple methodological perspectives.

Omics era

With gigabytes of new biological data emerging each day, filling database after database, it gets ever more clear that the conceptual and methodological backbone of the life sciences is currently undergoing dramatic changes. The *hypothesis-free collection* of data [45], e.g., the sequencing of whole genomes, and explorative data analysis have already trespassed into the domain of *hypothesis-driven research* [37], which has dominated the life sciences over the last centuries. Bold statements [9], that scientific reasoning will become obsolete in the near future and will be replaced by machine learning algorithms and correlation analyses, have been vividly debated in the community [26].

With all that data becoming publicly available, and biology becoming a data-rich science, it is no longer necessary nor even appropriate to conduct theoretical research for the sake of theory itself—a theoretical biologist's dream come true. Nonetheless, the search for natural laws and the development of rigorous theory, as opposed to the precise reproduction of observed behaviors using highly detailed mathematical model, should not be abandoned, as it has often been theoretical prediction and understanding that preceded major innovations in the natural sciences; and isn't it generally easier to ask questions about how things work rather than why things work the way they do? However, even in systems biology, there is still plenty of room for *hypothesis-driven* experimentation, and we will show that careful experimental design is necessary for the answering of particular questions (see Chapter 3).

Network biology

Networks consist of nodes, representing the entities of a system, and edges, describing the relations among them. This high level abstraction allows the description of basically any system in the universe. End of the 1990s, networks had a huge impact on biology and science in general, breathing new life into interdisciplinary research. Watts and Strogatz's seminal work on small-world networks and their properties [133] is a prominent representative of this hype, having been cited more than 4580 times[1] by now and ranking at #6 among the most cited papers in physics[2]. The thrilling possibilities of treating complex systems as diverse as metabolism [3], the air transportation system [48], or even the network of social interactions among super heroes [2], under a common framework of methods have attracted many researchers from different fields.

Generally, abstraction is a good thing. It allows us to arrive at decisions and take meaningful action, even in situations of overwhelming complexity. It lets us focus on

[1] According to ISI Web of Knowledge (Thomson Reuters)

[2] According to *Essential Science Indicators* (Thomson Reuters)

the critical points in our analysis, and most importantly, it lets us compare systems, which otherwise would not be comparable. Abstraction is the coordinated elimination of information from complicated situations. It can either facilitate the discussion of systemic properties—or impede it.

Building a metabolic network, i.e., a abstraction of a metabolic system, from a list of chemical reactions is straight forward: substrates are connected to reactions, and reactions are connected with their respective products. Following this procedure, one ends up with a bipartite network consisting of two disjoint sets of vertices, reactions r and metabolites m.

Of course, this bipartite representation can be collapsed into two distinct unipartite networks, either by projection onto r, leading to a reaction-centric representation, or by projection onto m, leading to a metabolite-centric representation of metabolism. In both cases information is lost [120], as the bipartite network cannot be reconstructed from the unipartite representations. But, whereas the degree distribution of the reaction-centric representation is unambiguous and unravels an additional piece of information, i.e., the number of reactions a reaction is connected to via its products, the metabolite-centric projection introduces a large number of cliques—all substrates of a reaction become connected to all products of the same reaction—biasing the connectivity of compounds participating in reactions with more than one substrate or product towards unnaturally high degrees. Furthermore, it is impossible to disentangle the number of reactions a metabolite participates from its metabolite-centric node degree. Thus, the metabolite-centric network representation of metabolism is a good example for an abstraction that can potentially lead to wrong assumptions about the system being discussed. However, the metabolite-centric representation has become the *de facto* standard for many studies and has been used in the majority of scientific treatments on metabolic networks [64, 131].

Additional metabolic network representations have been proposed in the past, including enzyme and gene-centric networks [61, 73, 117], and we will focus in our analyses mainly on the latter ones.

Steady-state and constraint-based modeling of metabolism

Evaluating the constraints that nature forces upon life provides an excellent framework for modeling biological processes under uncertainty [95]. Let

$$0 = \boldsymbol{S}\boldsymbol{v} \tag{1.1}$$

describe the steady state flux space of a metabolic system, where $\boldsymbol{S}$ denotes the stoichiometric matrix[3] of the system, $\boldsymbol{v}$ a vector of reaction fluxes, and 0 replaces the vector of the time derivatives of metabolite concentrations $\frac{d\boldsymbol{C}}{dt}$ usually found at this position. It is hereby noteworthy that eq. (1.1) is not a just a oversimplified approximation of metabolism, but under the right circumstances (see the introductory section) a valuable description of the system. Imposing upper and lower limits on the reaction fluxes, $v_{min} \leq v \leq v_{max}$, constrains the solution space to a finite size, which has a convex shape due to the linear nature of the system. Defining v_{min} and v_{max} allows one to incorporate (i) reaction directionality (setting $v_{min} = 0$ for irreversible reaction), (ii) experimentally measured fluxes [134], and (iii) medium conditions into the system.

A whole spectrum of steady-state and constraint-based modeling (CBM) methods have been developed over the last years in order to analyze the steady-state flux space described in eq. (1.1):

(i) Elementary mode analysis [119] and extreme pathways [92], two very similar concepts, enumerate all minimal reactions sets capable of steady state operation, i.e., all linearly independent $\boldsymbol{v}$ satisfying eq. (1.1). Extreme pathways basically constitute the basis solutions of the system, representing the edges of the convex polytope in a geometrical interpretation of eq. (1.1), and thus being a subset of elementary modes, which additionally include solutions on its surface and interior [107].

(ii) Because the methods described in (i) are intractable for genome-scale systems, Monte Carlo methods have been applied [122] in order to sample flux space.

(iii) Flux-coupling analysis [24] allows the determination of coupled reaction sets. The coupling type between to fluxes v_i and v_j can be determined by minimization and maximization of the following fractional optimization problem

$$R_{min}(R_{max}) = min(max)\ \{\frac{v_i}{v_j} : \boldsymbol{Sv} = 0, \boldsymbol{v_{min}} < \boldsymbol{v} < \boldsymbol{v_{max}}\}, \tag{1.2}$$

where $R_{min} = R_{max} = c$ implies full coupling (a non-zero flux for v_i implies a fixed flux for v_j), $R_{min} = c_1 > 0 \wedge R_{max} = c_2 < \infty$ with $c_1 \neq c_2$ constitutes partial coupling (a non-zero flux for v_i implies a variable flux for v_j), while $R_{min} = 0 \wedge R_{max} = c$ and $R_{min} = c \wedge R_{max} = \infty$ imply directional coupling (a non-zero flux for v_i implies a fixed flux for v_j but not necessarily the reverse). Fluxes v_i and v_j are uncoupled if $R_{min} = 0 \wedge R_{max} = \infty$.

(iv) Flux balance analysis (FBA), the most prominent method from CBM, uses linear programming to find a flux distribution that maximizes a specified objective [128], e.g., biomass or ATP production. It finds a solution to this optimization prob-

[3]The matrix $\boldsymbol{S}$ with dimensions $|m| \times |n|$ represents the stoichiometries of a reactions system containing metabolites m and reactions n, where $S_{i,j} < 0$, if reaction j consumes metabolite i, and $S_{i,j} > 0$, if it produces metabolite i.

lem

$$max\ \{\boldsymbol{c^T v} : \boldsymbol{Sv} = 0, \boldsymbol{v_{min}} < \boldsymbol{v} < \boldsymbol{v_{max}}\}, \tag{1.3}$$

where $\boldsymbol{c}$ denotes the coefficient vector of the target function.

(v) Recently flux balance analysis has been extended to incorporate experimental data that cannot be directly incorporated as flux boundaries, e.g, RNA transcript levels and proteomic data [18, 28, 112].

We will make use of the steady-state approximation of metabolism and apply CBM extensively throughout the thesis, for tasks like random media sampling, reaction deletion analysis, microarray data incorporation and assessment, as well as for flux-coupling analysis.

Thesis outline

This thesis is based on a cumulation of published and submitted work, as well as manuscripts being prepared for publication. In the following, each piece of work is presented in form of a chapter, put into a broader perspective by a concluding chapter. Short descriptions are given below, which explain the motivation and background of each work.

Chapter 2: Ranges of control in the transcriptional regulation of *Escherichia coli*

One of the most important open questions in genomics regards the spatial organization of genes on the genome. Are genes randomly distributed or do they follow some hidden organizational principle? And if so, is their spatial organization a major contributor to their transcriptional regulation? Transcriptional regulation in prokaryotes is mainly believed to be organized by a set of dedicated transcription factors, either down or up-regulating the transcription of their target genes, the *lac* operon [63] being one out of many examples of this regulatory principle. On the other hand, it is now generally accepted that chromatin plays a major role in eukaryotic gene regulation [75] and evidence is accumulating for this being true in prokaryotes as well [125].

RegulonDB [42], a database that collects information about transcriptional regulation in *Escherichia coli*, provided in its 6.2 version regulatory information for 1474 out of 4585 genes. Thus, regulatory information is missing for the vast majority of genes. So, why is this the case in one of the best studied model organisms? Of course, one can argue that nothing is ever known completely, but a time-resolved view on the evolution of the database shows that the number of regulated genes is stagnating over the last years (see Figure 1.1), indicating that perhaps for a significant portion

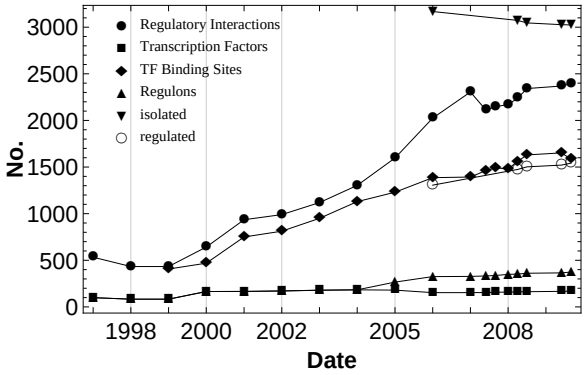


Figure 1.1: A summary of the amount of data provided by RegulonDB over the last 13 years.

of genes regulatory interactions and transcription factor binding sites will never be found.

In order to get some insights into this proposition, we applied, in collaboration with Dietrich and Helga Stoyan from *Bergakademie Freiberg*, point-process statistics to the spatial distribution of genes on the *E. coli* chromosome. In particular we investigated the distribution of two classes of genes: genes that are controlled by dedicated transcription factors and genes where no such information is available.

Chapter 3: Analog regulation of metabolic demand

Patterns of gene expression changes under variations of the underlying regulatory machinery are an important source of information for understanding the mechanisms of genetic control. We will show that the 3D structure of the circular chromosome of the model organism *E. coli* is one key component of this regulatory machinery. For this type of regulation, mediated by topological transitions of the chromosomal DNA, the term *analog control* was introduced by Marr et al. [81], in contrast to the regulatory action of transcription factors targeting specific DNA sites, denoted as *digital control* [81]. So far, the rich patterns of gene expression changes induced by alterations of the superhelical density of chromosomal DNA have been difficult to interpret.

In a collaboration with Georgi Muskhelishvili (Jacobs University) and Marcel Geertz (University of Geneva) we characterize effective networks formed by supercoiling induced gene expression changes, i.e., subgraphs which are exclusively composed of the dynamically active elements in the system. These effective networks are constructed by mapping the expression data onto static reconstructions of *E. coli's* metabolic and transcriptional regulatory networks.

Chapter 4: Multiple topological labels and medium-dependent essentiality in *Escherichia coli* metabolism

Based on previous well-known approaches to detect lethal reactions using topological and dynamical markers [8, 104, 136], resulting in reaction sets with surprisingly small overlaps, we explore the topological characteristics of reactions that are (i) always essential, (ii) essential only under specific media conditions, and (iii) never essential. Essentiality categories (i–iii) are determined *in silico* using random media sampling in conjunction with a single reaction knockout approach. For each sampled medium, the reaction targets for the knockout analysis can be determined by the identification of the set of active reactions. Other reactions do not have to be checked, as their removal will certainly not change the objective function optimum. The relative essentiality is than defined by the number of lethal outcomes of the removal of a target reaction divided by the number of environmental conditions this certain reaction was active. Furthermore, we investigate in a combinatorial fashion if combinations of the essentiality markers lead to more accurate essentiality predictions.

Chapter 5: A network perspective on metabolic inconsistency & Chapter 6: Metabolic variability in adrenal gland tumors

Adrenocortical adenomas are benign tumors; their medical significance lies not so much in their cancerous nature but rather in their ability to cause drastic imbalances in the endocrine system, e.g., primary aldosteronism [23], being characterized by an overproduction of the mineralocorticoid hormone aldosterone. Primary aldosteronism leads to a decreased renin activity and subsequent arterial hypertension.

In collaboration with Arndt Benecke and Annick Lesne, from *Institut des Hautes Études Scientifiques*, and Maria-Christina Zennaro, from *Institut National de la Santé et de la Recherche Médicale*, we integrate gene expression data obtained from healthy and tumorous adrenal glands with a genome-scale reconstruction of human metabolism [35] in order to understand the metabolic alterations in adenoma physiology. We also applied constraint-based modeling techniques to unravel physiological changes in the adrenal gland tumors, probing their abilities to produce energy as well as the hormone aldosterone. Furthermore, we compared our findings with results from a topological analysis (which we developed for the investigation described in Chapter 3), observing a strong agreement between both measures.

Chapter 7: Conclusions and outlook

This chapter combines and contextualizes the findings of the presented work and gives a comprehensive overview on the achieved knowledge gains. Furthermore, extensions to the presented work are proposed and a future outlook is given.

Chapter 2

Ranges of control in the transcriptional regulation of *Escherichia coli*

This chapter provides the content of the following publication [115]:

> Nikolaus Sonnenschein, Marc-Thorsten Hütt, Helga Stoyan, and Dietrich Stoyan *Ranges of control in the transcriptional regulation of Escherichia coli.* BMC Syst Biol (2009) vol. 3 (1) pp. 119

Abstract

Background: The positioning of genes in the genome is an important evolutionary degree of freedom for organizing gene regulation. Statistical properties of these distributions have been studied particularly in relation to the transcriptional regulatory network. The systematics of gene-gene distances then become important sources of information on the control, which different biological mechanisms exert on gene expression.

Results: Here we study a set of categories, which has to our knowledge not been analyzed before. We distinguish between genes that do not participate in the transcriptional regulatory network (i.e. that are according to current knowledge not producing transcription factors and do not possess binding sites for transcription factors in their regulatory region), and genes that via transcription factors either are regulated by or

regulate other genes. We find that the two types of genes ("isolated" and "regulatory" genes) show a clear statistical repulsion and have different ranges of correlations. In particular we find that isolated genes have a preference for shorter intergenic distances.

Conclusions: These findings support previous evidence from gene expression patterns for two distinct logical types of control, namely digital control (i.e. network-based control mediated by dedicated transcription factors) and analog control (i.e. control based on genome structure and mediated by neighborhood on the genome).

Background

The circular genome of *E. coli* is still an object of intense scientific research (see, e.g., [98]). It is a rich source of information on the organization of gene regulation, the interplay of different types of control exerted on gene expression, a model system for analyzing DNA topology, the model for which the most detailed electronically accessible transcriptional regulatory network has been compiled.

Many processes, acting on a broad range of scales, contribute to the evolution of bacterial chromosomes. Genes are organized in operons, i.e., groups of genes sharing a regulatory domain. The genome is shaped by point mutations, large-scale rearrangements, strand breaks and inversions during replication. The gene inventory is modified by gene duplications or deletions and lateral or horizontal transfer of genes. It is striking that an ever closer look at statistical properties of data reveals ever more systematic information, shaped by evolution, on an ever broader range of length scales.

Starting from the work by De Martelaere and Van Gool (1981) [82] and Jurka and Savageau (1985) [68] the gene density along the circular chromosome of E. coli has been discussed as a potential source of information on the evolutionary shaping of the system and in particular as a means of using DNA topology (i.e. the 3D structure of the genome) for regulatory purposes (see also [13]).

The papers by Warren and ten Wolde (2004) [132] and Képès (2004) [72] focus on distances between genes or operons. Both are studies of the specific patterns in the distributions of distances between regulatory pairs (genes or operons regulating each other or pairs of genes or operons co-regulated by other genes). Warren and ten Wolde (2004) [132] find a substantially reduced distance between operons in such regulatory pairs, suggesting an evolutionary pressure to reduce such distances for efficient regulation. For obtaining this, they use classical characteristics of point process statistics, namely partial pair correlation functions and nearest neighbor distance probability density functions.

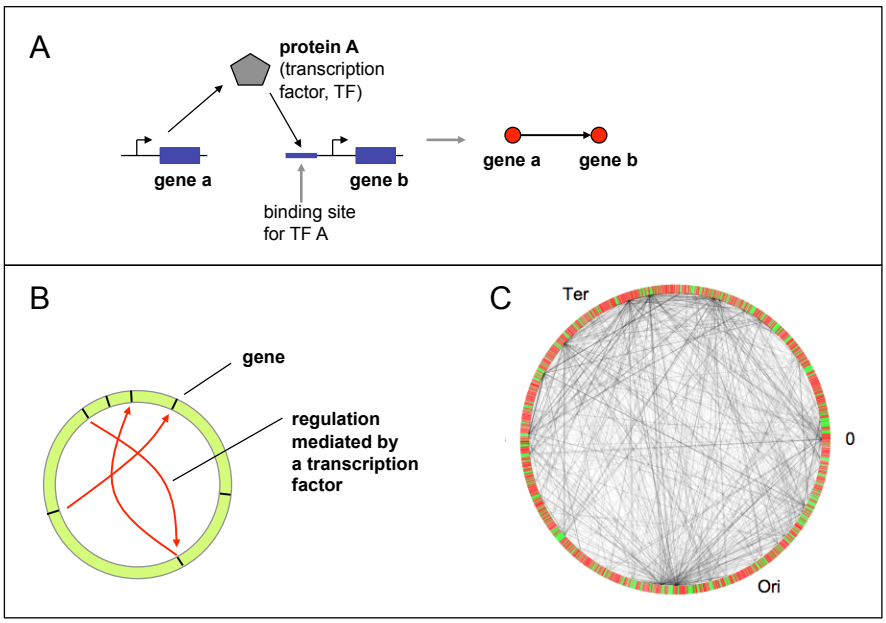


Figure 2.1: Schematic view on the transcriptional regulatory network. (A) For the TRN, the nodes are genes and the (directed) links describe the regulatory action of one gene onto another mediated by a transcription factor. More specifically, for the link shown in the Figure, gene a expresses protein A, which serves as a transcription factor (TF) binding in the regulatory region of gene b and thus controlling gene b. The links of the transcriptional regulatory network can be inserted into the circular genome of *E. coli* (schematically shown in (B) and for the real genome in (C), based on the data from RegulonDB, version 6.2.

Képès (2004) [72] observe a periodicity in the distances between regulator and target, where the period length is in the same order of magnitude as known loop domains in the 3D organization of the *E. coli* chromosome.

More recently, Hermsen et al. (2008) [56] observed that genes with opposite orientation have a bias towards larger distances, when oriented away from each other (divergent gene pair; e.g. the second gene pair in Figure 2.1) compared to those oriented towards each other (convergent gene pair; e.g. the first gene pair in Figure 2.1). They argue that this bias is due to the larger size of the upstream control region compared to the downstream control region.

Darling et al. (2008) [31] discuss biases in genomic inversions with respect to the replichores and other patterns of genome rearrangement in bacterial chromosomes. Another important factor influencing gene-gene distance statistics on a very general level is gene clustering. The origin of observed gene clustering is attributed to gene duplication and divergence, an evolutionary advantage of clustering, as it might in-

crease a gene's chance for horizontal gene transfer or, lastly, selective advantage of gene clusters due to functional coupling and the efficient organization of transcription (see the discussion in [38]).

From the systems perspective, mainly the regulatory control mediated by direct binding of transcription factors has been investigated. The compilation of these interactions for *E. coli* into a database [41] allows the construction of a transcriptional regulatory network (TRN) [109]. This view yields deep topological insights into the hierarchical organization of TRNs (Ma et al., 2004 [78]; Yu and Gerstein, 2006 [138]) and their composition out of specific network motifs (Shen-Orr et al., 2002 [109]). The TRN has been used for the interpretation of expression patterns (Gutierrez-Rios et al., 2003 [51]; Herrgard et al., 2003 [57]), revealing both the potential and the limitations of this perspective. In particular, recently it became obvious that other effects with very different regulatory mechanism have to be taken into account, like alterations of the DNA structure on a small [124, 125] and larger [135] scale. Thus, understanding the organizational logic of gene regulation necessitates a clear distinction of the different control types in the first place, as a prerequisite for the assessment of their impact in regulation.

Another link between these two research areas, gene distribution and TRN, comes from the observation that gene neighborhood explains some features of observed gene expression patterns (Marr et al. 2008 [81]; Blot et al. 2006 [20]). In particular, Marr et al. (2008) [81] analyze the interplay between two types of control in gene expression profiles in *E. coli*, one network-mediated and the other mediated by DNA topology.

These two control types have been termed *digital* (referring to the fact that the TRN provides static information on the connections between unique, discontinuous components, e.g. a particular pair of regulator and regulated gene) and *analog* (referring to the fact that the expression of specific genes is under the control of continuous information provided by distributions of supercoiling energy in the genome), respectively [81].

The statistical properties of gene distributions and gene spacings have been studied to detect deviations from randomness and interpret these deviations in a suitable evolutionary context. To a large extent, these investigations differ (apart from the technical details of the statistical tools and the construction of suitable null models) predominantly in the categories of genes analyzed. In the present paper we show results for two analysis steps, where the first analysis distinguishes between two classes. Analysis I discusses genes involved in regulation (i.e., either being regulated by a transcription factor or producing a transcription factor regulating other genes; class 1) and genes not involved in regulation mediated by transcription factors (which in the following we will call "isolated genes"; class 2). Analysis II consists of pairs of genes regulated by a common transcription factor. Distances between the genes in such a pair will be contrasted to the distances between arbitrary genes.

The biological hypothesis behind these categories is that different means of gene regulation essentially have different length scales. The novel feature of our approach lies in two points: (1) the distribution of regulated/regulating genes vs. (regulatorily) isolated genes has not been studied before. Our finding here, a pronounced deviation from randomness for the isolated genes, fits to the hypothesis stemming from previous investigations of control types in gene expression patterns (Marr et al. 2008 [81]); (2) in order to detect deviations from randomness we employ different non-classical types of correlation functions.

Our hypothesis, based on the findings from Marr et al. (2008) [81], is that the existence of distinct logical types of control (namely digital and analog) has a systematic impact on the statistical features of gene distributions. In particular, distances between isolated genes and all others should be smaller than average distances between genes, as isolated genes tend to be co-regulated by spatial neighborhood via the 3D structure of the genome.

Results are in the following presented both on the level of individual genes and on the level of operons.

Results and Discussion

First we present the gene distance distributions for the two gene classes, (isolated genes and genes involved in regulation; see above). Then we discuss pair correlation functions $g(s)$, partial pair correlation functions $g_{ij}(s)$, mark connection functions $p_{ij}(s)$, connectivity correlation functions $c(s)$, and control correlation functions $k_3(s)$ (see Materials and Methods).

Figure 2.2 explains the categories of genes (operons) we are studying. In the first part (Analysis I; classes 1 and 2; cf. Figure 2.2A and C) we are looking at statistical properties of shortest distances between two genes involved in regulation (s_{11}), two isolated genes (s_{22}), and an isolated gene, together with a gene involved in regulation (s_{12}); cf. Figure 2.2B. In the second part (Analysis II) we study distances between two genes regulated by a common transcription factor. In all cases we analyze the shortest distance (in base pairs, bp) along the circular genome to the nearest neighbor of the respective type. We do not consider the orientation of genes on the genome or the sizes of the genes and operons. In fact, we represent every gene only by a single point, namely by its center. We checked that our results do not change qualitatively when we consider other definitions of the "distance" between two genes (e.g., from start points to end points or the minimal distance between any two points of the two genes; cf. Figure 2.2D). We ignore biases induced by the relative position of the gene under consideration with respect to the origin of replication (ori) or the Ter macrodomain, respectively.

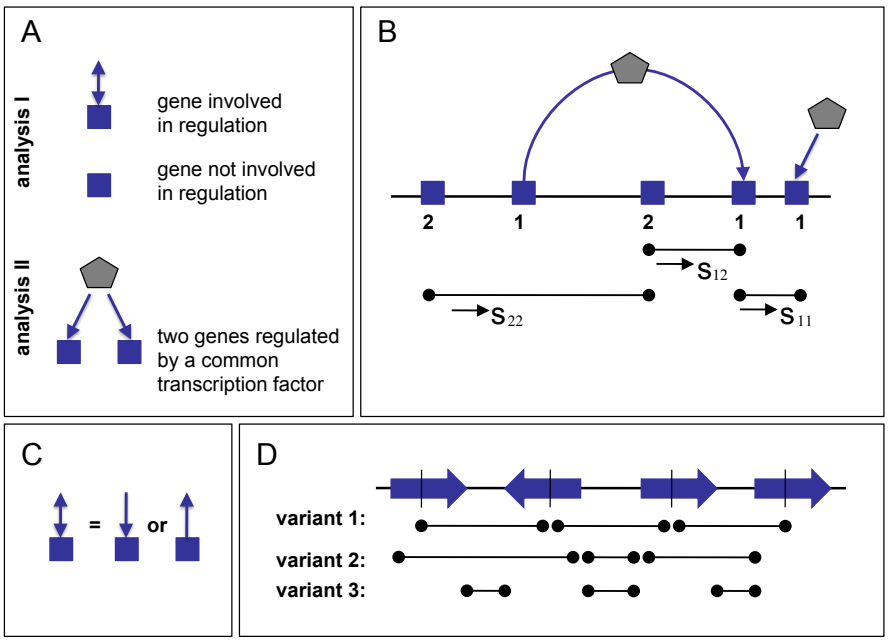


Figure 2.2: Definitions of categories and distances. (A) The classes of genes (involved and not involved in regulation) entering Analysis I and the subset of genes involved in regulation (pairs of genes under common regulation), which is studied in Analysis II. (B) Examples of distances s_{11}, s_{12} and s_{22} entering point process analysis. (C) No distinction is made between genes receiving a regulatory influence and genes encoding transcription factors regulating other genes. (D) Various possibilities of defining the distance of two genes along the genome. In this investigation we use variant 1.

Thus we are confronted with problems of point process statistics (see Materials and Methods), where the genes or operons are the points. They are marked by 1 or 2, corresponding to the classes above, 1 : involved in regulation, 2 : isolated.

Figure 2.3 shows the distributions of shortest distances on the gene level (Figure 2.3a) and on the operon level (Figure 2.3b), together with the distribution of gene content of operons (Figure 2.3c; i.e. the number of genes in an operon). Most of the operons in *E. coli* consist of only a single gene, some operons, however, contain as many as 15 genes. Figure 2.3 already reveals several interesting features of the data: Distances between operons tend to be larger than distances between genes (which results from the systematic omission of intra-operon distances, when passing from genes to operons); the decrease in frequency with the distance does not seem to follow an exponential distribution, suggesting a deviation from a Poisson process (and therefore from a random distribution of points).

After the discussion of the nearest neighbor distances we report now on the correlation functions. First we discuss the pair correlation functions $g(s)$, see Figure 2.4.

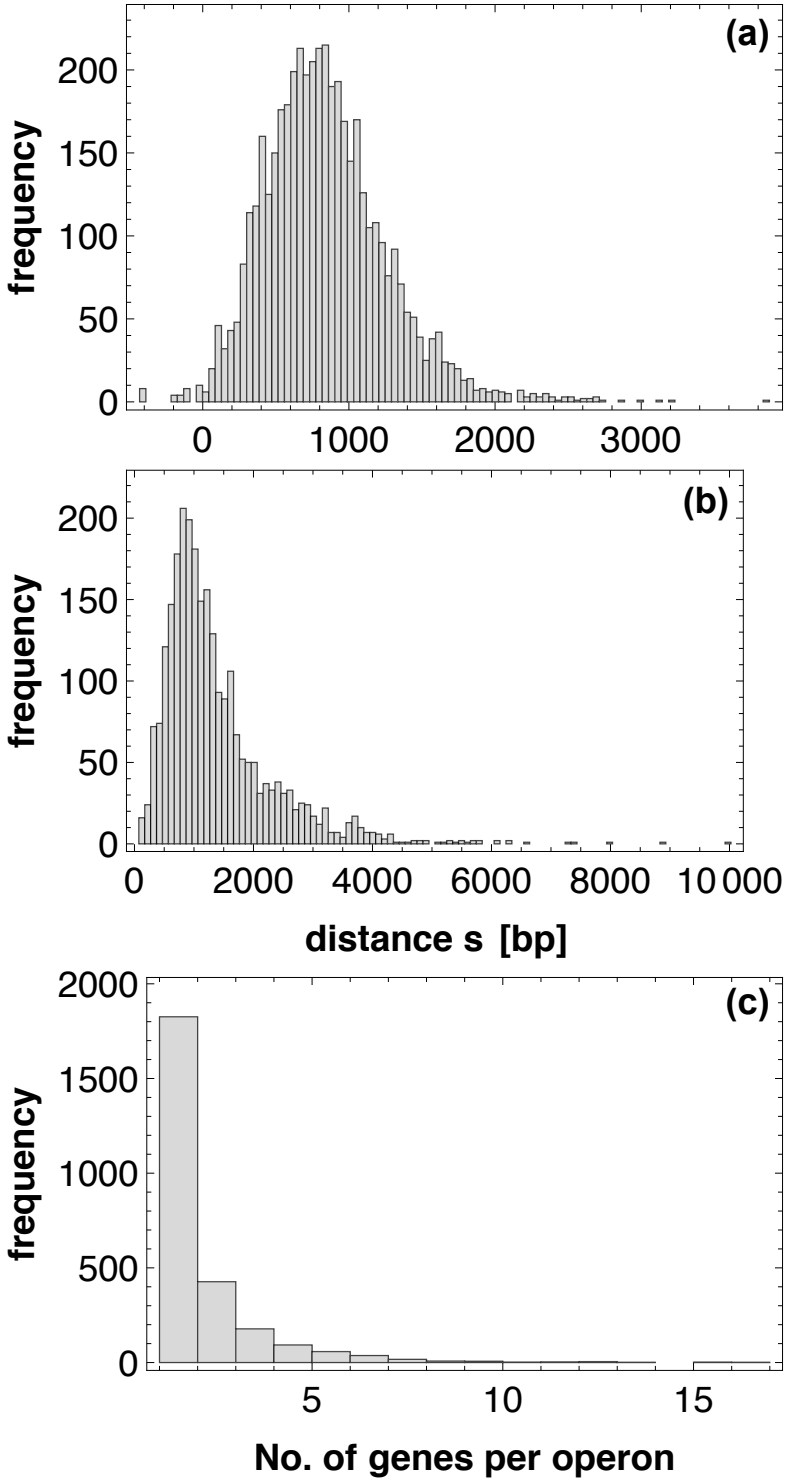


Figure 2.3: Pairwise distances and operon sizes. Histogram of pairwise distances for (a) genes and (b) operons, and distribution of operon sizes (c).

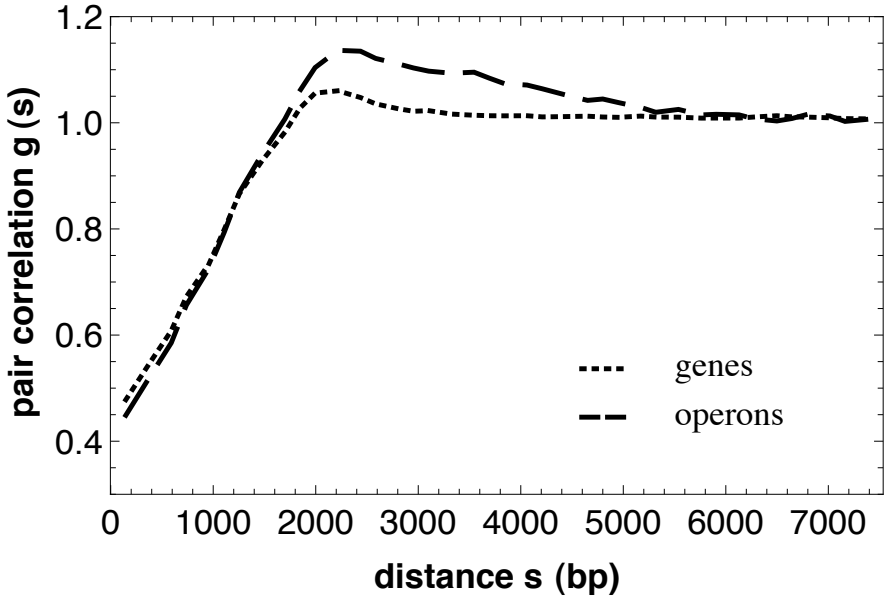


Figure 2.4: Pair correlation functions. The pair correlation function $g(s)$ for genes (dotted) and operons (dashed). The functions indicate a weak tendency of regularity of the gene/operon positions.

They indicate the well-known fact that the positions both of the genes and operons are not completely randomly distributed, i. e. according to a Poisson process, where $g(s) = 1$. In contrast, they are more regularly distributed, probably simply due to the finite size of the objects, as is shown by the values of $g(s)$ smaller than 1 for small s.

Typical gene sizes range from a few hundred bp to several thousand bp with the mean size centered around 1 kpb.

There is even a maximum for distances of 1 and 2 kbp, while for larger values the curves approach fast the value 1, which corresponds to absence of location correlation. The range of correlation is for the operons somewhat longer than for the genes, it goes until 6 kbp.

A suitable tool for analyzing the relative contributions of the different categories to these correlations is the partial *pair correlation function* $g_{ij}(s)$ with $ij = 11$ (between genes involved in regulation), $ij = 22$ (between isolated genes) and $ij = 12$ (one gene involved in regulation, the other isolated), respectively. Figure 2.5 shows the curves $g_{ij}(s)$ for genes (Figure 2.5a) and for operons (Figure 2.5b). The results for g_{11} and g_{22} are similar to those from Warren and ten Wolde (2004) [132] and, in fact, display similar features as the *pair correlation function* $g(s)$ in Figure 2.4: very small distances are suppressed (due to the finite size of the elements); one observes a peak between 1 and 2 kbp and then a convergence to the value 1 as for the uncorrelated case. The ranges of correlation are between 5 kbp and 7 kbp.

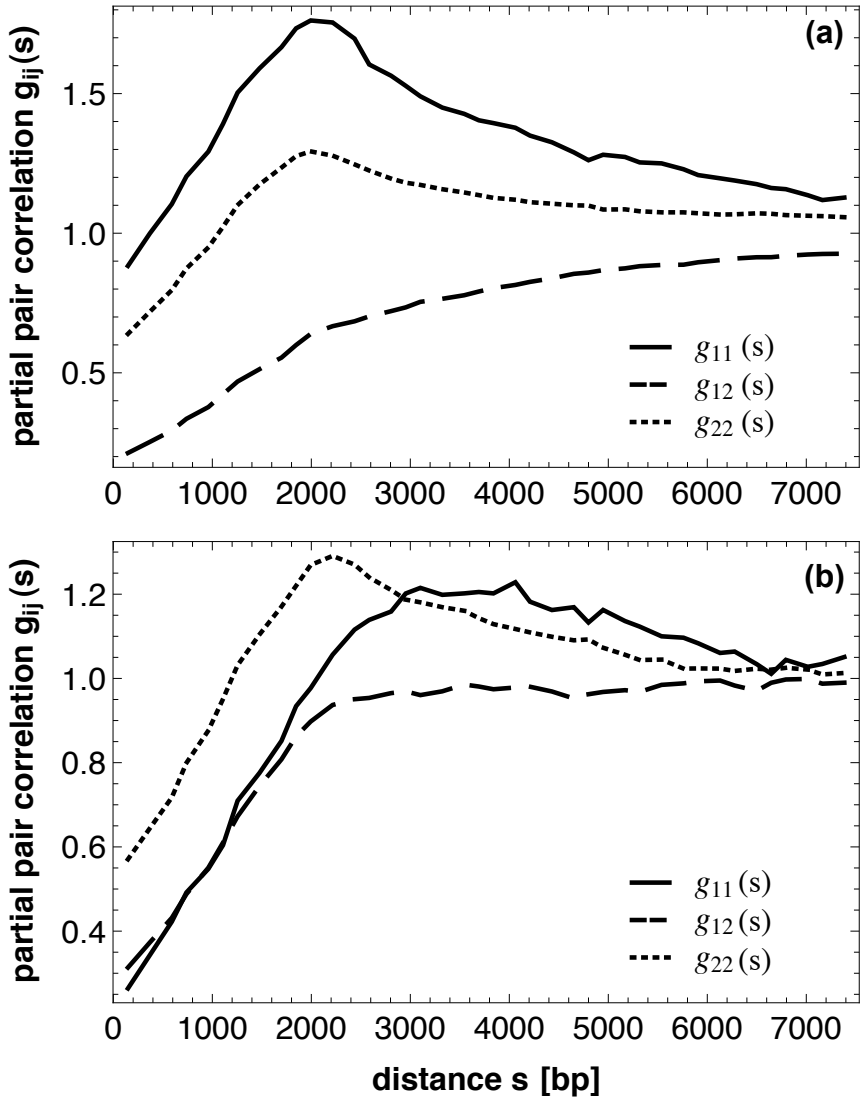


Figure 2.5: Partial pair correlation functions. Partial pair correlation functions $g_{ij}(s)$ for (a) genes and (b) operons. In both cases the full curve denotes $g_{11}(s)$, the dotted curve $g_{22}(s)$ and the dashed curve $g_{12}(s)$. The functions indicate a weak tendency of regularity of isolated and regulated points, and a clear tendency of repulsion between isolated and regulated points.

The curves for $g_{12}(s)$, however, are new and, to a certain extent, unexpected: the distances between isolated and regulatory genes do not show a peak at intermediate distances. Obviously, the repulsion between isolated and regulatory genes is stronger and longer than that of genes of the same type, namely 7 kbp. In contrast, for operons it is shorter, only 3 kbp.

The term "repulsion" is used here in a simplifying sense, in order to say that there is a tendency that the distances between isolated and regulatory genes are larger than between genes of the same type. This may be a result of real repulsion as well as of relative "attraction" of the members of one class towards itself.

We interpret this repulsion as an unmixing of genes predominantly regulated by transcription factors (digital control; cf. [81]) and genes predominantly regulated by the 3D structure of the genome (analog control). For the first type (class 1) distance correlations should be less important than for the second type (class 2) where regulation is mediated (among other processes) by the neighborhood of genes on the genome.

Figure 2.6 shows the *mark connection functions* $p_{ij}(s)$, again for genes (Figure 2.6a) and for operons (Figure 2.6b). All these function are simply monotonous and nearly linear. The curve for $p_{22}(s)$, for example, shows that the probability that two genes (operons) of distance s are both isolated is monotonously decreasing. The numerical differences between the values for genes and operons result from the different values of p_2, which are 0.676 and 0.735 for genes and operons, respectively. These functions show that the maxima of the $g_{ii}(s)$ result mainly from higher numbers of gene/operon pairs of the corresponding inter-gene/operon distances, while the probabilities that members of such pairs are both monotonously decreasing in s. They are decreasing for $i = 2$ for genes and operons, while the function $p_{11}(s)$ is decreasing for genes and increasing for operons. The decrease of $p_{22}(s)$ in both cases (genes and operons) is in good agreement with our expectations that short distances should contribute more strongly in the case of analog control. We believe that the decrease of $p_{11}(s)$ for genes is a consequence of the very strong short-range contributions of intra-operon distances (i.e. of genes within the same operon).

It should be noted that the *partial pair correlation functions* $g_{ij}(s)$ compared to the *mark connection functions* $p_{ij}(s)$ are individually normalized. In contrast to $p_{ii}(s)$ we see maxima of $g_{ii}(s)$ around 2 kbp. Comparison between the types 1 and 2 shows that regulatory genes are more regularly distributed than isolated genes (as the maximum is higher for $g_{11}(s)$). We would also like to point out that the estimates of the partial pair correlation function and mark connection function depend continuously on the proportions of class 1 and class 2 genes in this analysis (see also Methods). We thus expect that small fluctuations in the data will leave the main results of our analysis intact.

Both, in the *partial pair correlation functions* $g_{12}(s)$ and in the *mark connection function* $p_{12}(s)$ one can see that the two classes (isolated genes and genes involved in

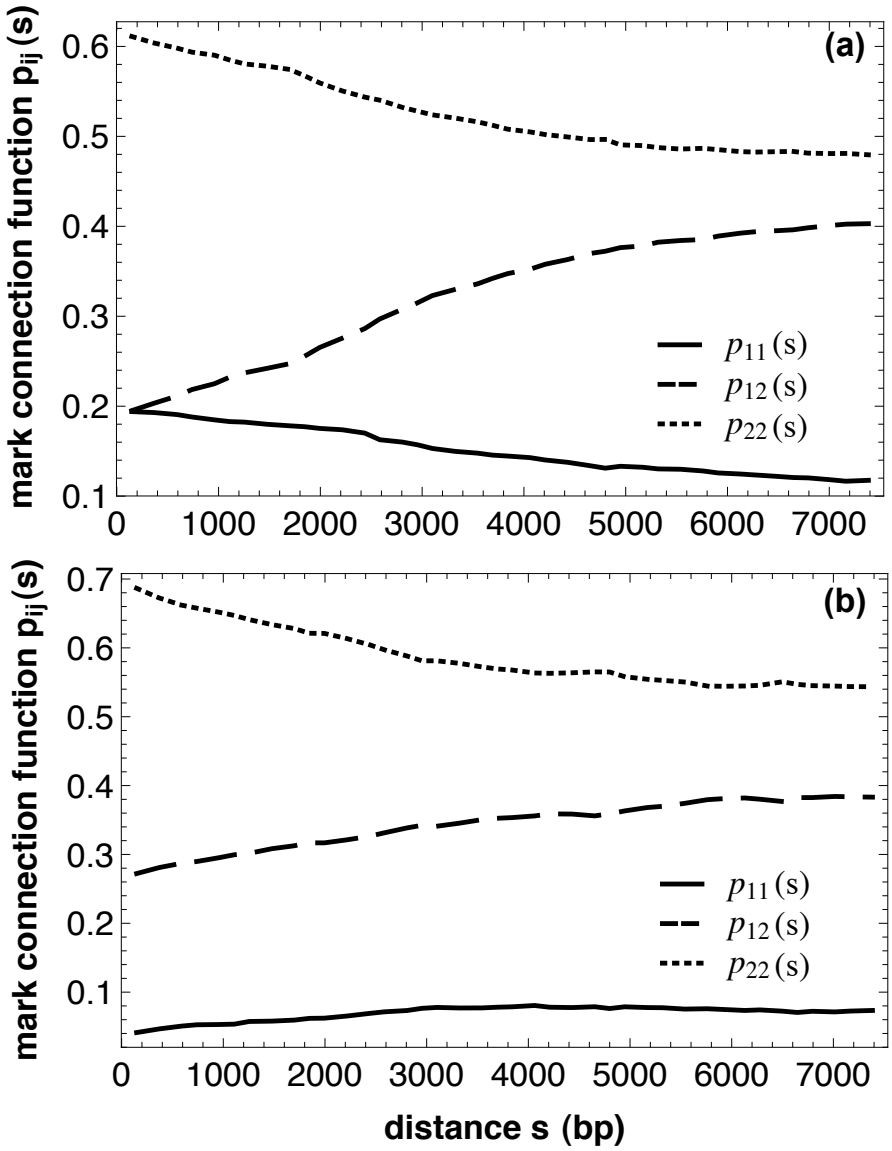


Figure 2.6: Mark connection functions. Mark connection function $p_{ij}(s)$ for (a) genes and (b) operons. Analogously as in Figure 2.5, the full curves denote $p_{11}(s)$, the dotted curves $p_{22}(s)$ and the dashed curves $p_{12}(s)$. The functions show that the maxima in Figure 2.5 result only from different frequencies of inter-point distances, while the probabilities of being isolated or regulated depend monotonously on the interpoint distance s.

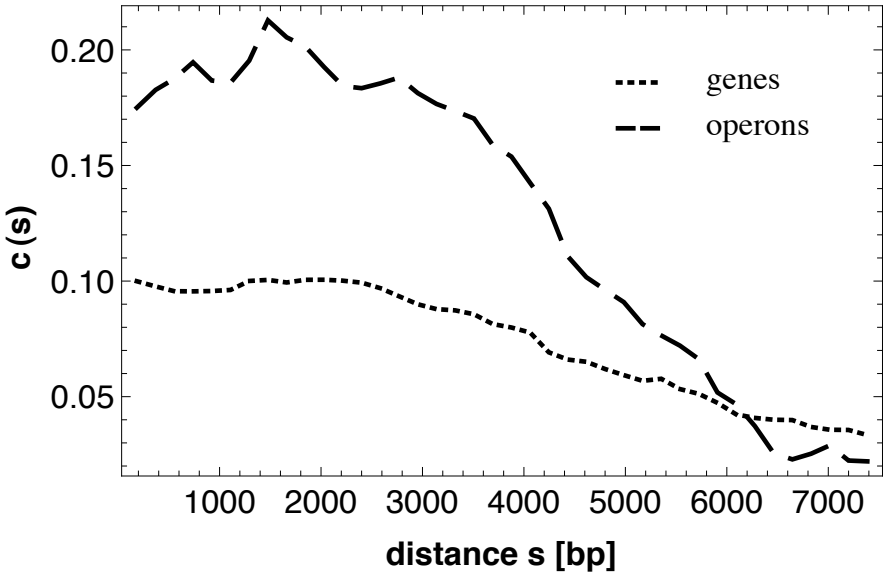


Figure 2.7: Connectivity correlation function. Connectivity correlation function $c(s)$ for genes (dotted) and operons (dashed). The functions show that the probability that the members of a pair of non-isolated points regulate each other decreases only for distances s larger than a value s_0 (approximately 3000 bp for the genes and 2500 for the operons). The weak irregularities of the curve for the operons result from the fact that, while the same estimator is used as for the genes, the number of passive operons is much smaller than that of that of passive genes and so the statistical quality of the results decreases a little.

regulation) repel each other. On the level of the operons this repulsion is less clearly visible (and has a range up to approximately 2.5 kbp); in general, operons are more irregularly spaced than the genes. In all these cases, this can be explained by the elimination of many short (intra-operon) distances from consideration, when passing from the gene level of description to the operon level.

The second group of correlation functions describes distances between points (genes/operons) under common regulation. First we look at the probability for two regulatory (mark 1) points at a distance s that there is a regulation relationship between them. This is expressed in terms of the connectivity correlation function $c(s)$, which is given in Figure 2.7. The curves show that there is a critical distance s_0 between 2 and 3 kbp such that for s larger than s_0 the probability that the two points regulate another decreases continuously. The numerical values for the operons are clearly larger than those for the genes, indicating a higher systematic (importance of distance for the organization of regulation).

Finally, Figure 2.8 shows the curves for the control correlation functions $k_3(s)$. Now only passive points are considered, a subset of the regulatory points. The prob-

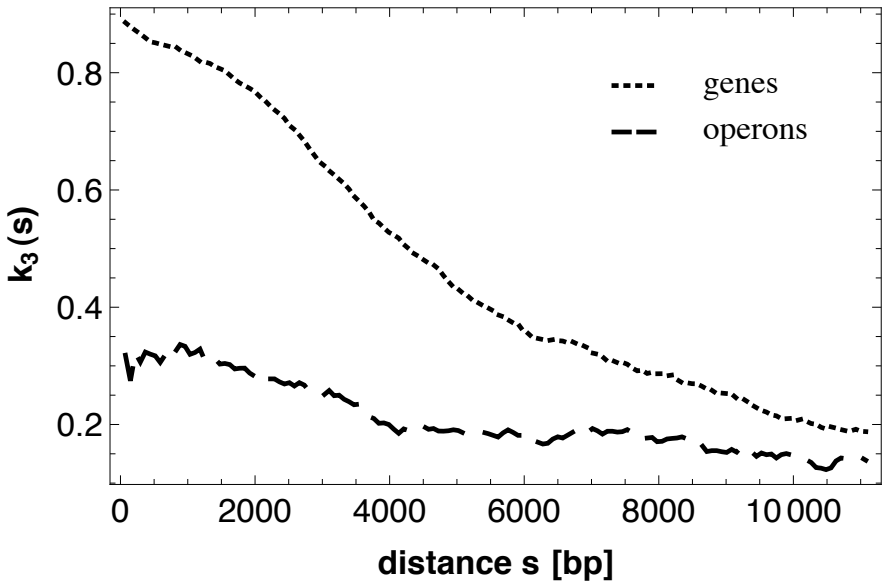


Figure 2.8: Control correlation function. Control correlation function $k_3(s)$ for genes (dotted) and operons (dashed). The functions show that the probability that the members of a pair of passive points are regulated by the same point decreases monotonously with increasing distance s. As in Figure 2.7, the weak irregularities of the curve for the operons result from the fact that, while the same estimator is used as for the genes, the number of passive operons is much smaller than that of that of passive genes and so the statistical quality of the results decreases a little.

ability of interest is that the two points considered are regulated by the same (active) other point. Since the basic point processes for $c(s)$ and $k_3(s)$ are different, namely 1-points and passive points, no simple inequality between both functions must hold true.

The function $k_3(s)$ thus indicates that the three objects involved (two regulated genes, one regulator) are preferentially close together.

How do the two categories, regulatory genes (class 1) and isolated genes (class 2), compare with experimental information on gene regulation? We used a list of supercoiling-sensitive genes from [94] and compared it with the two categories of genes discussed in our manuscript (*i*: number of isolated genes, and *r*: number of genes involved in regulation). Figure 2.9 shows the ratio of isolated and regulatory genes for both experimental classes (*s*: number of supercoiling-sensitive genes; *n*: number of non-sensitive genes; *si* then denotes the number of isolated supercoiling-sensitive genes, etc.), normalized by the number of genes in the two categories; for clarity, a value of one (representing equal proportions of genes in both categories) has been subtracted. The first value in Figure 2.9 is thus $((si/sr)\ (r/i) - 1)$. If we assume that supercoiling-sensitive genes are genes, for which analog control is systematically

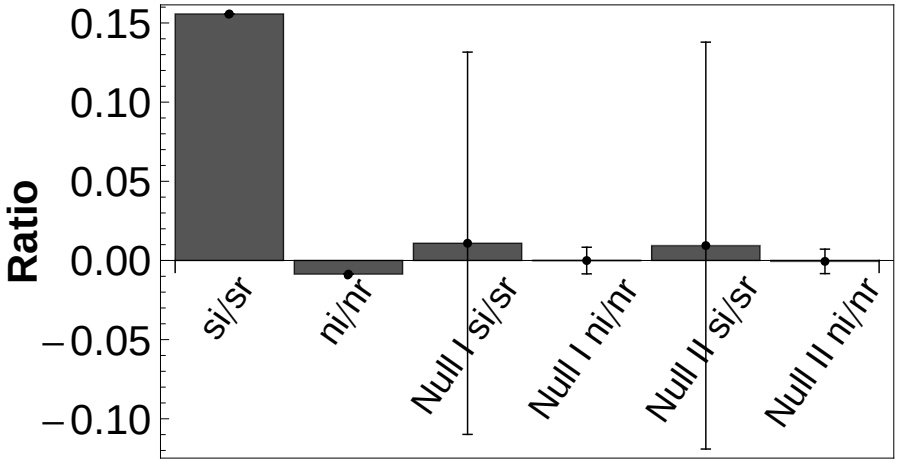


Figure 2.9: Comparison with supercoiling-sensitive genes. Excess of isolated vs. regulatory genes in the supercoiling-sensitive genes, together with a comparison with randomly drawn genes (null model I: randomly selected supercoiling-sensitive genes; null model II: randomly selected isolated genes).

more important than digital control, we expect a larger percentage of isolated genes to be in this group. Even though this test can only provide very indirect evidence, the effect is clearly visible in Figure 2.9, as the first value deviates the strongest from zero. For non-sensitive genes, as well as for all random samplings the values are close to zero. It should be pointed out, however, that the statistical significance is not high enough to form a solid basis for interpretation. The more sophisticated techniques, which lead to the previous figures, are indeed necessary for this.

Our statement that short distances and analog control are qualitatively related can also be checked on the level of this data set. While it should be noted that our key result is a statistical signal emerging from the collective ensemble of genes (and here we show additionally, how these findings can again be cross-validated against high-throughput data), we again resort to the data from [94] and compare a histogram of inter-gene distances obtained from supercoiling-sensitive genes with a histogram obtained from a random selection of genes. The trend towards smaller distances is clearly seen. This figure is included as supplementary information (Additional File 2.10).

The second data set is taken from [130], where the protein occupancy landscape (i.e. the probability per base of a protein binding event) has been measured. The authors distinguish between transcriptionally silenced extensive protein occupancy domains (tsEPOD) and highly expressed extensive protein occupancy domains (heEPOD). Figure 2.11 shows the distance of isolated genes and regulatory genes from these domains in a cumulative plot. While no substantial difference between the isolated

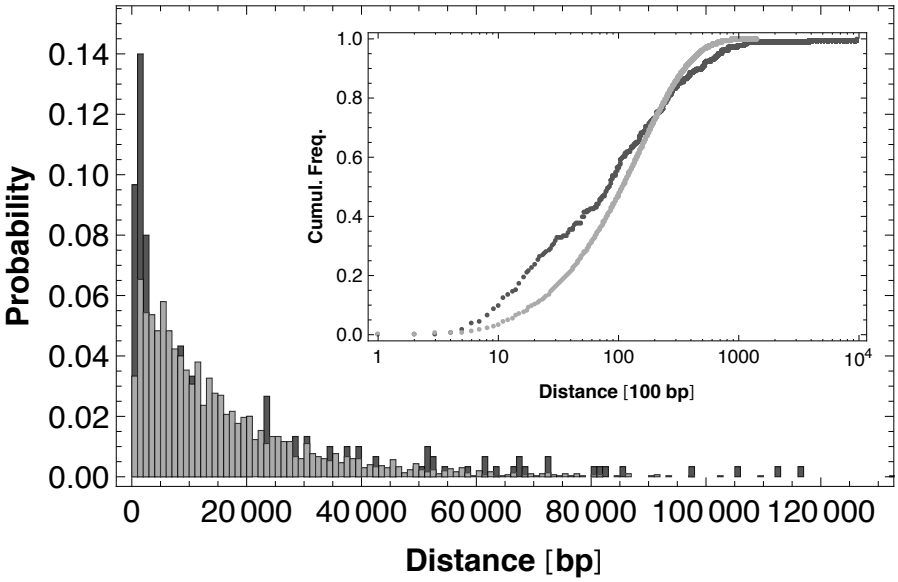


Figure 2.10: Distances among supercoiling-sensitive genes and other genes. Histogram of distances observed between supercoiling-sensitive genes (dark gray) and a random sample of other genes (light gray). The inset shows the corresponding cumulative distance plot.

and regulatory genes is seen for the heEPODs (Figure 2.11b), the distances of isolated genes to tsEPODs are clearly shorter than those of regulatory genes (Figure 2.11a), pointing again towards a biological significance of this distinction between isolated and regulatory genes and also towards a stronger importance of analog control (mediated by regional binding events of structural proteins to the DNA) for isolated genes.

The inset in Figure 2.11b summarizes the two parts of Figure 2.11 by showing the difference between the isolated gene curve and the regulatory gene curve from Figure 2.11a (full curve in the inset; tsEPODs) and from Figure 2.11b (dotted curve in the inset; heEPODs), respectively. A particular interesting feature seen in the inset is that at short distances the full curve goes up and the dotted curve goes down, i.e. there are (at short distances) far more isolated genes in the vicinity of transcriptionally silenced EPODs and more regulatory genes in the vicinity of highly expressed EPODs.

Lastly, we looked at pairs (class 2, class 1) = (i, r) of genes, taking into account the orientation of genes on the respective strand. Figure 2.12 distinguishes between $i - r$ and $r - i$ and shows the cumulative distances for these two cases ($r - i$: full, $i - r$: dotted). We find strong differences between these cases, again suggesting that "isolated" and "regulatory" are meaningful categories for our analysis. Additionally, these differences could indicate that the larger size of the regulatory regions in the r category is a contributor to the repulsion we observe between the two categories.

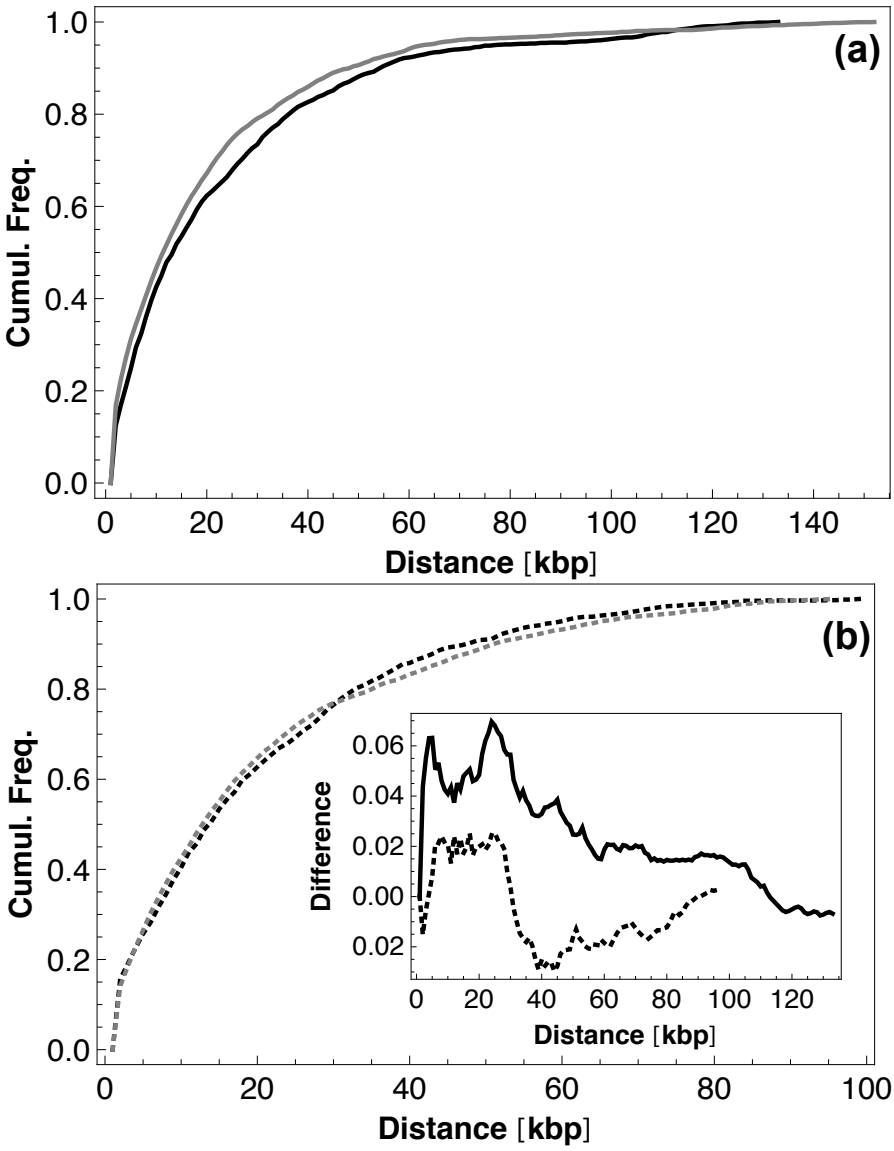


Figure 2.11: Comparison with extensive protein occupancy domains (EPODs). Cumulative frequencies of distances between genes and extensive protein occupancy domains (EPODs): (a) distances to transcriptionally silenced EPODs (full curves), (b) distances to highly expressed EPODs (dotted curves). In both cases this analysis has been performed independently for the isolated genes (gray curves) and the regulatory genes (black curves). Inset: Differences between the gray and black curves from both parts, i.e. for tsEPODs (full curve in inset) and heEPODs (dotted curve in inset).

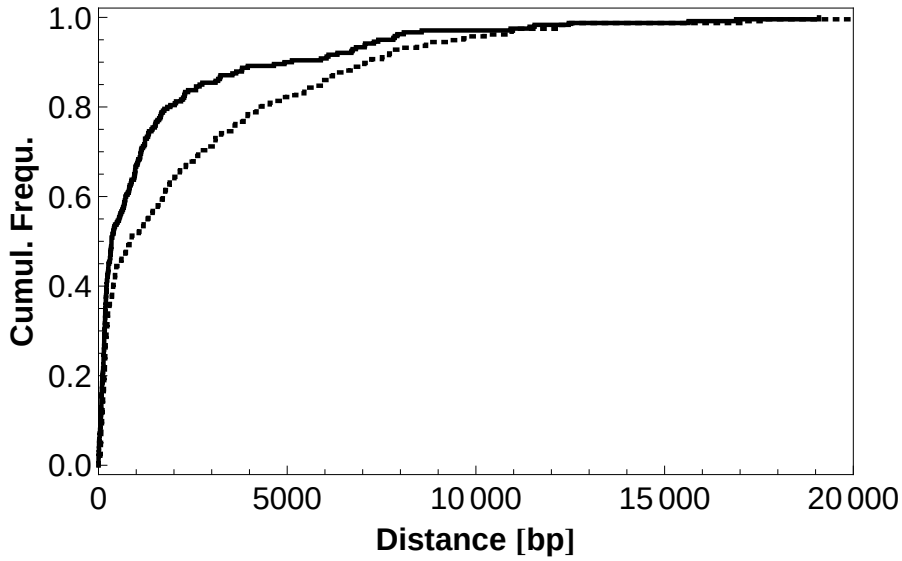


Figure 2.12: Distances between regulatory and isolated genes taking orientation into account.

Conclusions

Patterns (i.e. systematic deviations from randomness in the arrangement of genes) in the genome of *E. coli* have been studied on many different scales.

Here we analyzed another facet of this topic by distinguishing between genes involved and not involved in regulation based on transcription factors. Our key finding is that these two classes, regulatory and isolated genes display a statistical repulsion. Furthermore, the (operon-level) partial pair correlation function has a peak at shorter distances for isolated genes than for regulatory genes. This preference of shorter distances for isolated genes is also visible in the mark connection function and is supportive of our hypothesis that analog control is more important for this class of genes than for the regulatory genes, for which digital control is a longer-ranging alternative.

Whether the statistical properties of inter-gene distances discussed here originate from the need to organize gene regulation or from the dynamics of genome rearrangement cannot be ultimately decided based on the data at hand.

Minimal models of genome arrangement dynamics and its impact on gene expression could be a useful tool for deciding whether the distance pattern between genes is indirectly shaped (and therefore deviates from pure randomness) by these dynamics, rather than being evolutionarily constraint to contribute more directly to gene regulation.

The statistical differences between isolated and regulatory genes described here suggest that, indeed, the genes currently classified as isolated from the perspective of

the available TRN are systematically different from the genes involved in regulation. We by no means want to suggest that (a) these genes are indeed unregulated nor (b) that the current version of RegulonDB (version 6.2) is complete. However, when considering the extreme cases of isolated genes being just gaps in the database and, on the other hand, isolated genes being systematically regulated by other means, our results support the latter view.

Even though we consider our findings in an evolutionary context (by making visible some deviations from randomness of the gene distances in the *E. coli* genome, which can only be understood evolutionarily) we here do not directly discuss the comparative genomics aspect of it. It would be particularly interesting to analyze the degree of evolutionary conservation as a function of the distance between genes and separately for the two categories of genes. A hypothesis for such an extension of our analysis could be that pairs of genes contributing strongly to the patterns we observe, have a higher degree of evolutionary conservation. This is, indeed, a whole work package we plan to tackle in a future investigation.

Eventually one needs to arrive at a more holistic view of the system and explain the interplay between gene arrangement, DNA binding site distributions, physical properties of DNA binding sites, the architectural properties of the transcriptional regulatory network and the spatial gene expression patterns, in order to understand the binding site code behind global gene expression and to unravel the universal design principles of transcriptional regulation.

Methods

Point process statistics

In the statistical analyses of this paper, the genes are considered as points on a circle C, the circular chromosome of *E. coli*. Thus, a random system of points is analysed, which leads to the application of methods of point process statistics. (The term "process" is related to early applications where the points were time instants. Also the term "stationary" is related to these applications; "homogeneous" could be an equivalent.) These methods have been mainly developed for the planar $(d = 2)$ and spatial $(d = 3)$ case, but can be easily applied also in the one-dimensional $(d = 1)$ case considered here. So our main reference is Illian et al. (2008) [62].

Similarly to the investigations of [72, 132] we assume that the point pattern belongs to a "stationary" point process, i. e. that the point distribution is rotation invariant. This implies that the local point density does show only irregular fluctuations, as it is the case. Thus it makes sense to speak about the "intensity", the mean number of points per length unit. As in [62] it is denoted here by λ.

The points considered are marked. There are two marks, namely "1" and "2", where "1" stands for "regulatory" and "2" for "isolated". The fraction of i-points is denoted by p_i, for $i = 1, 2$. Note that p_i can be interpreted as the probability that a randomly chosen point has the mark i. Furthermore, the probabilities p_1 and p_2 make sense, where p_i is the fraction of i-points (a point with mark i) in the point process. It can be interpreted as the probability that a randomly chosen point is an i-point.

The statistical analysis uses a series of summary characteristics, which have been successfully employed in spatial point process statistics. All these functions depend on a variable s, which is a distance. In all cases this is the shortest distance along the circular genome.

All these function can be called "correlation functions", but not all include only point pairs; therefore some of them are not second-order characteristics.

The best known function is the *pair correlation function* $g(s)$, which is explained here as in [62], p. 219, since the explanation there is closer to the "two-point interpretation" used for explaining the other functions. (The explanation in Warren and ten Wolde is different but equivalent.)

Consider two points x and y on C of distance s and two infinitesimal length elements of lengths dx and dy centred at x and y. Denote the probability that in the two elements there is each a point by $p(x, y)$. This probability is given by the so-called product density $\varrho(x, y)$ as $p(x, y) = \varrho(x, y)dxdy$.

In the stationary case (which is assumed), $\varrho(x, y)$ depends only on the distance s of x and y, and the simpler symbol $\varrho(s)$ is used. The pair correlation function is then $g(s) = \varrho(s)/\lambda^2$.

The normalisation by division by λ^2 makes that for large r, $g(r)$ approximates 1. Values of $g(r)$ larger than 1 for small r indicate clustering, while values smaller than 1 indicate some tendency of regularity or repulsion between the points. See the discussion of the information given by a pair correlation function in [62], pp.219.

The *partial pair correlation functions* $g_{ij}(s)$ are defined using refined product densities $\varrho_{ij}(s)$ where one of the points in the infinitesimal intervals is an i-point and the other a j-point, see [62], p. 325. These functions are normalized by $p_i p_j \lambda^2$, which leads to $g_{11}(s)$, $g_{12}(s)$ and $g_{22}(s)$. The $g_\alpha(s)$ in [132] are similar to $g_{11}(s)$.

Again, the normalisation leads to values around 1 for large s, and also the general interpretation is similar to that of $g(r)$, see [62], pp.325. For $i \neq j$ the relations between different sorts of points are characterized. For example, values smaller than 1 for $g_{ij}(r)$ indicate some tendency of repulsion or inhibition between points of the different types i and j.

The *mark connection functions* $p_{ij}(s)$ are defined by

$$p_{ij}(s) = \frac{\varrho_{ij}(s)}{\varrho(s)}, \tag{2.1}$$

of course only for such s where the denominator is positive, see [62], p. 331. It can be interpreted as the conditional probability that two points at distance s have marks i and j, given that these points are in the point process. These probabilities have the following behavior for large s:

$$\lim_{s\to\infty} p_{ii}(s) = p_i^2 \tag{2.2}$$

and

$$\lim_{s\to\infty} p_{ij}(s) = 2p_ip_j \tag{2.3}$$

for $i \neq j$. It is useful to consider the *mark connection functions* additionally to the partial pair correlation functions since they characterize the occurrence of the point types with eliminated influence of fluctuations in point density; see [62], p. 332.

Comparison of the Figures 2.5 and 2.6 shows the power of this approach. The curves in Figure 2.5 are heavily dominated by the frequencies of point distances, which show for the genes a maximum at around $s = 2000...3000$, while Figure 2.6 shows the true nature of the marking: the probability that two points of distance s have, for example, both mark 2 decreases monotonically with s.

The connectivity correlation function $c(s)$ is also a characteristic of a conditional nature. It is defined by

$$c(s) = \frac{\varrho_{conn}(s)}{\varrho_{11}(s)} \tag{2.4}$$

where $\varrho_{conn}(s)$ is a quantity which yields the probability that between the points x and y in the infinitesimal intervals above, if both are regulatory (both have mark 1), there is a direct regulatory relationship, i. e. one of them regulates the other or both regulate the other. It is similar to the connectivity function in [62], p. 249, and can be interpreted as the conditional probability that between two regulatory points at distance s there is a direct regulatory relationship.

Finally, the control correlation function $k_3(s)$ is defined by

$$k_3(s) = \frac{\varrho_3(s)}{\varrho_{pp}(s)}. \tag{2.5}$$

It is defined for the sub-point process of all points that are regulated by other points ("passive" points, a subset of all 1-points); its product density is denoted by $\varrho_{pp}(s)$. Furthermore, $\varrho_3(s)$ is a quantity which yields the probability that for two passive points x and y in the infinitesimal intervals above there exists a third point which regulates both x and y. Thus, $k_3(s)$ can be interpreted as the conditional probability that for two passive points at distance s there is a third point which controls both of them.

Transcriptional regulatory network and spatial distribution of genes

We obtained the data from RegulonDB (version 6.2) [41], which is a database specifically dedicated to the transcriptional regulation of *E. coli*. A total number of 4548 genes are included in this database, of which 1474 bear information about their transcriptional regulation and thus have been classified as class 2 genes.

Authors contributions

DS and MH conceived the study. HS, DS and NS analyzed the data. DS, HS, NS and MH wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

MTH acknowledges support by Volkswagen Foundation. NS is supported by a Jacobs University scholarship. We are indebted to Georgi Muskhelishvili (Bremen, Germany) and Carsten Marr (Munich, Germany) for helpful comments on the manuscript.

Chapter 3

Analog regulation of metabolic demand

This chapter provides the content of the following manuscript submitted for publication [117]:

> Nikolaus Sonnenschein, Marcel Geertz, Georgi Muskhelishvili, and Marc-Thorsten Hütt *Analog regulation of metabolic demand*, *submitted*

Abstract

Background: The 3D structure of the chromosome of the model organism *Escherichia coli* is one key component of its gene regulatory machinery. This type of regulation mediated by topological transitions of the chromosomal DNA can be thought of as an *analog* control, complementing the *digital* control, i.e. the network of regulation mediated by dedicated transcription factors. Here we show that DNA supercoiling, in cooperation with the nucleoid associated proteins (NAPs), i.e. the analog level of control, coordinates gene expression with metabolism. We analyze the pattern of gene expression changes induced by alterations of the superhelical density of chromosomal DNA from a network perspective.

Results: We find a significantly higher correspondence between gene expression and metabolism for the wild type expression changes compared to mutants in NAPs, indicating that supercoiling induces meaningful metabolic adjustments. As soon as

the underlying regulatory machinery is impeded (as for the NAP mutants), this coherence between expression changes and the metabolic network is substantially reduced. This effect is even more pronounced, when we compute a wild type metabolic flux distribution using flux balance analysis and restrict our analysis to active reactions. Furthermore, we are able to show that the regulatory control of DNA supercoiling is not mediated by the transcriptional regulatory network (TRN), as the consistency of the expression changes with the TRN logic of activation and suppression is strongly reduced in the wild type in comparison to the mutants.

Conclusions: So far, the rich patterns of gene expression changes induced by alterations of the superhelical density of chromosomal DNA have been difficult to interpret. Here we characterize the effective networks formed by supercoiling-induced gene expression changes mapped onto reconstructions of *E. coli's* metabolic and transcriptional regulatory network. Our results show that DNA supercoiling coordinates gene expression with metabolism. Furthermore, this control is acting directly because we can exclude the potential role of the TRN as a mediator.

Background

A single *Escherichia coli* chromosome comprises 4.6 Mb and must be compacted at least ~10^3 fold to fit inside the bacterial cell. Despite tremendous compaction the nucleoid is a dynamic structure adapted to varying rates of replication and different transcriptional requirements resulting from changes in environmental conditions. This double requirement of compaction and differential gene expression implies that bacterial chromatin must possess a high degree of spatial organisation. Recent investigations indicate that the maintenance and utilisation of negative supercoils in the DNA is central to both issues [124].

In the protein-free DNA molecule, DNA superhelicity is partitioned into a twist component, Tw, which is reflected in a twisting or untwisting of the double helix for positively and negatively supercoiled DNA respectively, and a writhe component, Wr, which is a measure of the three-dimensional path of the double helical axis. In a closed topological domain these quantities are related to a change in linking number (ΔLk) from the relaxed state such that $\Delta Lk = \Delta Tw + Wr$. Negative supercoiling can facilitate both DNA folding and compaction as well as the untwisting of DNA which is required for the initiation of transcription and replication. The introduction of negative supercoils into DNA requires energy that is then available for driving processes such as transcription and replication initiation. This energy is described by the equation $E = k_B T \Delta Lk^2 / 2\langle \Delta Lk^2 \rangle$, where $\langle \Delta Lk^2 \rangle$ represents the variance of the Gaussian distribution of DNA topoisomers. Importantly the available energy is proportional to the square of the difference in linking number. Since the total link-

ing number is a constant for a closed domain, twist and writhe can be partitioned in different ways and hence the DNA can assume different structures [84].

Gene promoter regions are generally characterised by high deformability, being susceptible to duplex destabilisation under conditions of superhelical stress [55, 87, 123]. The cellular promoters can be thus understood as devices channeling the free energy of negative supercoiling to localised, biologically relevant sites in DNA. Several studies using different promoters and promoter derivatives revealed that there is a distinct, yet characteristic, coupling between the superhelical density of DNA and the activity of a particular promoter [11, 99, 106]. A change of supercoiling could thus globally and differentially affect the efficiency of channeling superhelical energy at distinct promoters, allowing coordinated change of gene expression activities to occur.

Besides classical modes of transcriptional regulation through dedicated transcription factors (the transcriptional regulatory network), which we would like to refer to as *digital control* [81], it is well known that DNA topology affects gene expression in prokaryotes [20] as well as in eukaryotes [75], which we call *analog control* ([81]; see Figure 3.1A).

In the bacterial cell the abundant nucleoid associated proteins (NAPs), including FIS, H-NS, HU, Lrp, Dps and IHF, fulfill the role of packaging and dynamic constraint of superhelicity. These NAPs are assumed to be mediators of analog control exerted by long-range nucleoprotein structures formed by binding of multiple low affinity sites in the chromosome as opposed to digital control exerted by low concentrations of dedicated transcription factors binding specific DNA sites with high affinity [81].

In particular, this combination of a global state (i.e. the superhelical density) and local states (domains and chromatin) is responsible for the spatial transcript patterns observed along the chromosome [20, 53, 65, 94].

DNA supercoiling is homeostatically controlled by topoisomerase I and DNA gyrase [114]. Furthermore, the superhelical density is responsive to a range of physiological conditions, e.g. the growth phase ([14]; see also Figure 3.1B), phosphorylation potential of the cell [127] and stress conditions [25]. It is precisely this physiological dependence that prompted us to ask whether DNA supercoiling is a global regulator relating the chromatin structure and transcription to metabolic demand [20, 85].

In order to answer this question on a system level we utilize alterations of superhelical density to measure supercoiling induced gene expression changes together with a combination of NAP mutations (FIS and H-NS, see Figure 3.1C and 3.2), thus precluding the buffering effects of the homeostatic network [106].

We are here discussing the interpretational capacity of the cell: the environmental information is sensed by and filtered via chromatin structure. We show that the regulation of the metabolic state is predominantly achieved by this analog type of control.

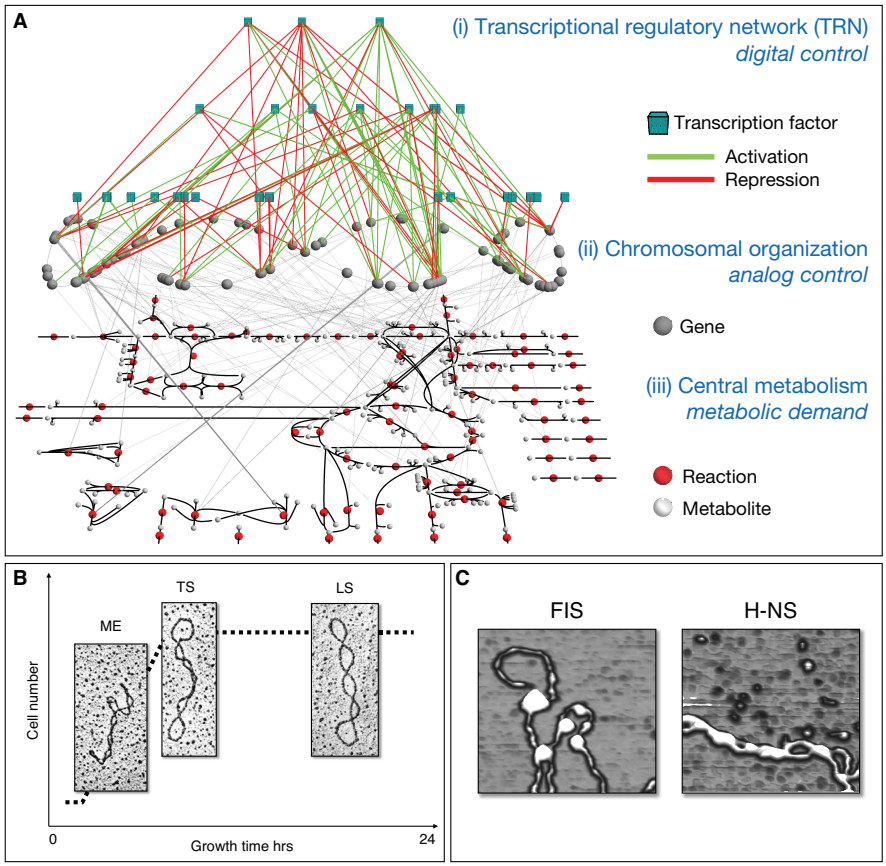


Figure 3.1: Illustration of the different components involved in *E. coli* transcriptional regulation, transcription and metabolism.

Furthermore, we show that the regulatory control of DNA supercoiling is not mediated by the transcriptional regulatory network (TRN), as the consistency of the expression changes with the TRN logic of activation and suppression is strongly reduced in the wild type in comparison to the mutants. Our data are evidence for an optimal conversion of supercoiling into metabolic adjustments by NAPs.

While it is true for eukaryotes that the multi-level organization of gene regulation obfuscates the connection between mRNA and protein levels, let alone metabolic fluxes, and it seems that most of the control on metabolism is contributed by the post-transcriptional levels [30], the situation is known to be quite different in prokaryotes where transcription and translation are tightly coupled [46, 139]. So it is valid to analyze the role of transcriptional regulation in order to understand bacterial homeostasis and the metabolic state of a cell.

To our knowledge, this work is first to show directly on a system-wide level the coordinated regulation of cellular metabolism by DNA supercoiling and NAPs.

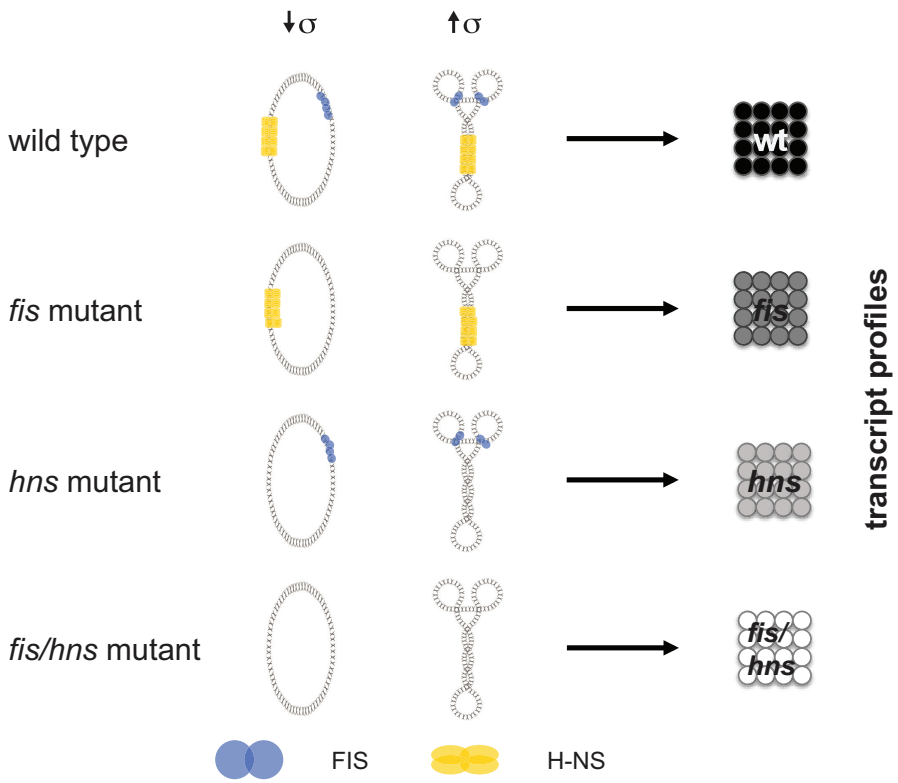


Figure 3.2: Experimental setup. Transcript profiles of four *E. coli* strains (wild type, *fis* mutant, *hns* mutant and *fis*/*hns* double mutant) are compared under low ($\downarrow \sigma$) and high superhelical density ($\uparrow \sigma$). The shading of the schematically depicted data sets on the right-hand side (black, dark gray, gray and white) will be used throughout the article.

Results and Discussion

Analysis strategy

An important feature of our approach is that we analyze subnetworks of the overall metabolic gene network defined by the data at hand. These *effective networks* contain only the active components (differentially expressed genes) under the given conditions (alterations in the superhelical density) and are analyzed from a network-topological perspective. The connectivity of these gene-centric effective networks is thus a result of the underlying reaction-centric topology, together with the observed gene expression pattern. Deviation of this connectivity from randomness is what we will in the following call *metabolic coherence* (MC). A second, more refined definition of effective metabolic gene networks, which will also be used in the following, requires both a significant expression change for one of the associated genes and a non-zero metabolic flux predicted for the encoded reaction using flux balance analy-

sis [128] under specified environmental conditions.

The coherence of metabolism and gene expression patterns is quantified as follows (details are given in *Material and Methods* and Text A): we map the patterns of differentially expressed genes from the four genetic backgrounds (wild type; *fis*, *hns*, *fis*/*hns* double mutants, respectively) directly onto a metabolic gene network in order to extract effective networks. Then we compute the ratio of connected nodes and all nodes in the effective network, which we call *metabolic coherence ratio* (MCR). This quantity is then converted into a z-score, by using a random distribution of expression changes as a null model (Figure 3.3 summarizes this procedure), which is our *metabolic coherence* (MC) in the following. The MC allows us to compare the amount of network coherence between gene expression profiles and metabolic pathways for the different data listed in Figure 3.2.

In order to validate our results on a broad scale we use network reconstructions from multiple independent databases and also apply different methods to handle gene-reaction mappings as well as currency metabolites (see also *Materials and Methods* and Text A). In the following we will present our results for the different variants of the metabolic coherence for the four gene expression profiles from Figure 3.2.

Metabolic coherence

In Figure 3.4, the four values of the MC (for the wild type and the three mutants) are shown for three different metabolic network representations, namely for the EcoCyc database [70], for the KEGG database [69] and for the *i*AF1260 metabolic model [39].

Figure 3.4A displays the pattern retrieved from the gene network based on the EcoCyc pathways. The wild type expression data exhibit the strongest coherence with the metabolic network (high MC). The wild type also shows the strongest MC for the KEGG network compared to the three mutants, however less clearly than for the EcoCyc case (see Figure 3.4B). Figure 3.4C gives the MC pattern for the *i*AF1260 gene network, from which we manually removed currency metabolites. In this *in silico* model of *E. coli* metabolism, we again observe a strong MC for the wild type and low values for the mutants, with the double mutant exhibiting the lowest amount of coherence.

Flux balance analysis (FBA) as a quantitative approach for simulating steady-state fluxes on metabolic networks [128] allows us to study, whether the observed systematics are enhanced, when only active links in the metabolic network are taken into account. Using the *i*AF1260 model, we computed a steady-state flux distribution that maximizes biomass production [39] under a rich medium condition and eliminated all inactive links from the network. The resulting MC is shown in Figure 3.4D. Strikingly, the restriction to active fluxes enhances the previous pattern (from Figure 3.4C) for the metabolic coherence.

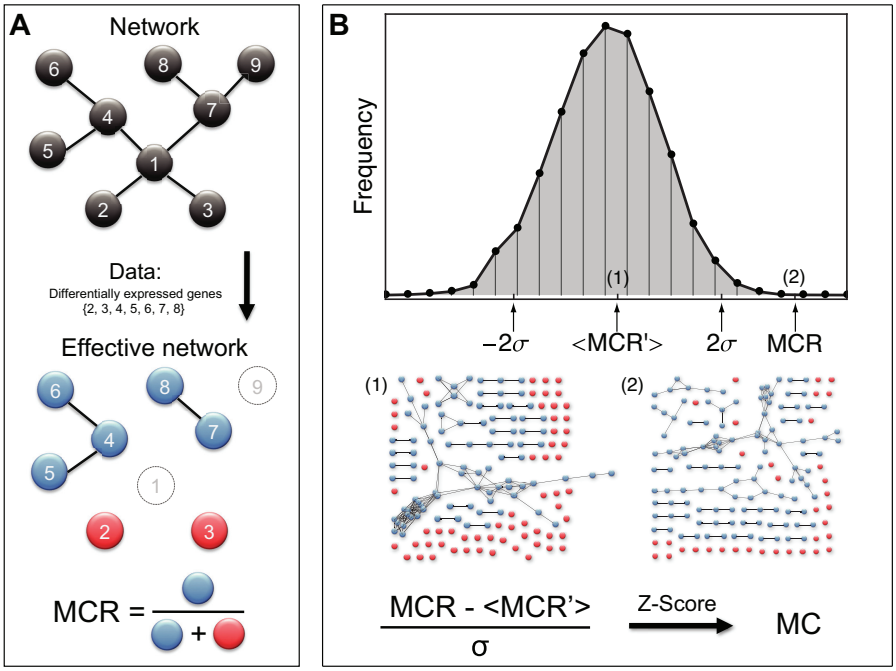


Figure 3.3: Effective gene networks and metabolic coherence. (A) Scheme depicting the calculation of the metabolic coherence ratio MCR for a fictitious network and data set. The data (genes with significantly changed expression) are mapped onto the network resulting in an effective subnetwork. The metabolic control ratio is then the ratio of connected nodes (blue) and all nodes, i.e. the sum of connected and isolated (red) nodes, in the effective network. Unhighlighted nodes correspond to genes with no significant expression changes. (B) Calculation of the metabolic coherence. Randomly reselecting the same number of affected nodes in the network allows the sampling of random effective networks and thus the computation of a set of random metabolic control ratios MCR'. These allow the computation of a z-score value termed metabolic coherence MC for the MCR. (2) is an example of a real effective network, whereas (1) is one of its random counterparts. MCR' of (1) lies approximately around the mean $< MCR' >$.

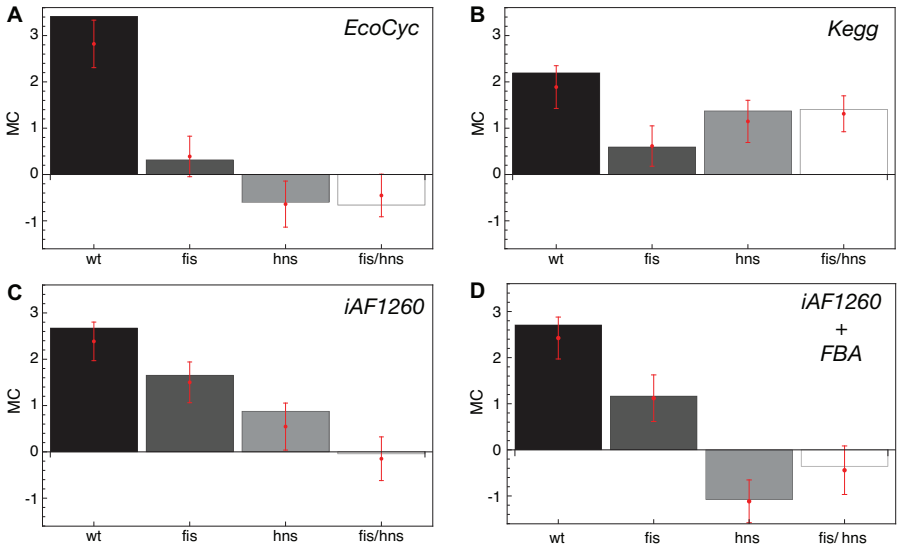


Figure 3.4: MC for four independent *E. coli* metabolic network reconstructions. (A) The network obtained from the EcoCyc database pathway information. (B) The network obtained from the KEGG pathways. (C) The network subset of the *i*AF1260 network where currency metabolites have been removed manually. (D) The *i*AF1260 network (currency metabolites have been removed manually) consisting only of reactions active under a rich medium condition. Error bars represent the standard deviation of a jackknife test, where the MC was recomputed 100 times by discarding 10 % of the transcript data for each of the four genetic backgrounds.

The key observation from Figure 3.4 so far is that changes in gene expression levels brought about by changes in supercoiling energy in the genome have a strong metabolic interpretation: the agreement of these expression changes with the metabolic network is significantly above randomness (as measured by the metabolic coherence). When severely perturbing the internal mechanisms of chromatin organization (by eliminating FIS and/or H-NS from the system), metabolic coherence goes down.

Robustness of the result

Network analysis has established itself as an efficient way of exploring biological systems ([47, 81]; see also Text A). Nevertheless, network treatment of metabolic systems is accompanied by certain difficulties and we check the robustness of our results against many of them. In order to solidify this initial result, we need to look in detail at several issues, which can potentially affect our analysis (see also Text A):

(*i*) Gene to reaction mapping. While all our analyses have been performed with gene-centric graphs, the reaction-centric graph serves as the starting point for assessing metabolic information (in particular, the activity of metabolic fluxes). Decisions

are therefore necessary, how to relate the reaction level with the gene level. The procedure of mapping genes (i.e. the layer of information, where expression changes occur) onto reactions (i.e. the layer of information, where the metabolic network is evaluated) can have an impact on our result.

(*ii*) Treatment of currency metabolites. Currency metabolites are compounds in metabolic reactions balancing charge, energy, phosphate etc. They are distinguished from main metabolites (which define the metabolic pathway structures) only by biochemical knowledge or, qualitatively and indirectly, due to their very high degree in the metabolic network (resulting from their involvement in a vast number of reactions). The treatment of currency metabolites is an important issue in the discussion of the topological properties of metabolic networks (see, e.g., [77]). An approximate way of eliminating currency metabolites from metabolic network representations is to remove a certain percentage of highest-degree metabolites. Alternatively, one can use a database, where metabolites are already labeled as main metabolites and currency metabolites, respectively. This information is included in the most recent variants of the KEGG database (e.g., release 51.0; see [69]). In the *E. coli* FBA model *i*AF1260 [39], this information is not available. In order to obtain a currency metabolite free version of *i*AF1260 we used either a threshold to remove 4 % of the most highly connected metabolites (*threshold heuristic*; comparable to the procedure described in [73]) or a manually curated network (resembling the procedure described in [77]; see also Text A).

(*iii*) Differences between metabolic databases. Using intersections of the different metabolic reconstructions of *E. coli* allows us to focus on the commonalities between them.

(*iv*) Definition of the growth medium for determining the active metabolic reactions via FBA.

All these points are addressed in the following.

Large-scale evaluation

Figure 3.5 shows the MC signatures (sorted by size of the wild type MC) for a large compendium of metabolic gene networks. These networks can be subdivided into five categories (MC values for all networks and data sets shown in Figure 3.5 can be found in Table A.1):

(*i*) The most basic setup are metabolic gene networks extracted from the EcoCyc, KEGG and *i*AF1260 database.

(*ii*) In order to evaluate the influence of the gene-reaction mapping on our results, we computed MC values for all databases using the following configuration: (a) Taking all multiplicities into account, (b) excluding cases where a single or multiple genes are associated with two consecutive reactions and (c) taking only reaction

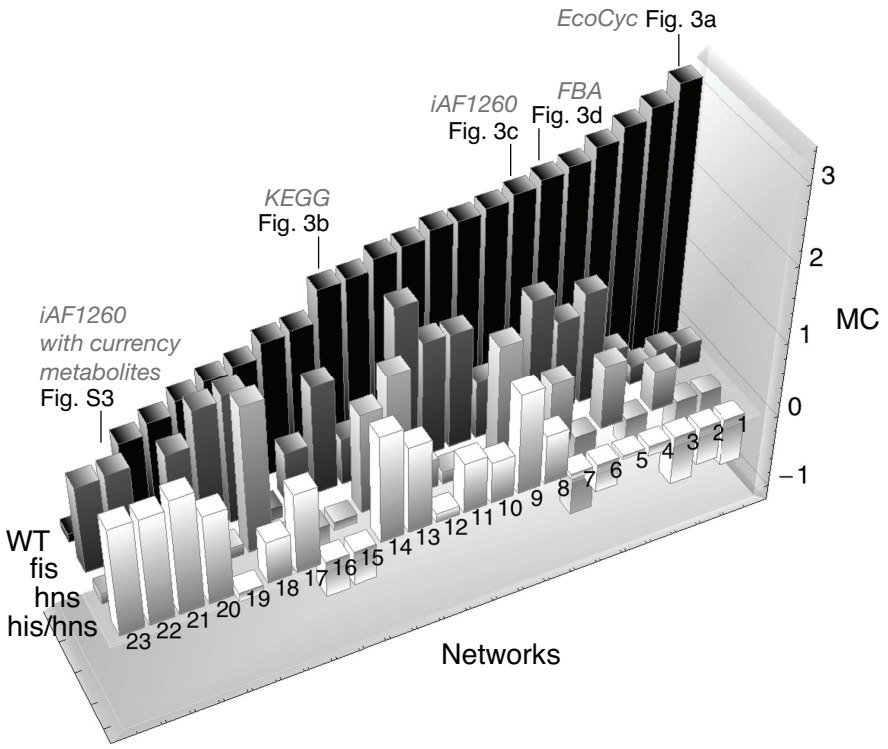


Figure 3.5: Results for the MC analysis for all available network reconstructions sorted by the size of the wild type MC. Notation: *network** – linked reactions with an overlap in the underlying gene set have been omitted; *network*** – only linked reactions are included, where both are associated with single non-overlapping genes; iAF1260$_{man}$ – currency metabolites have been removed manually; iAF1260$_{deg}$ – currency metabolites have been removed by degree threshold; iAF1260 – the untreated network (KEGG and EcoCyc are per construction free of currency metabolites); in the following (k) denotes slice number k in the chart. (1) EcoCyc, (2) EcoCyc*, (3) Intersection of EcoCyc and KEGG networks, (4) Intersection of EcoCyc and iAF1260$_{man}$, (5) iAF1260$^*_{man}$, (6) iAF1260$_{man}$ obtained from FBA (rich medium), (7) iAF1260$_{man}$, (8) EcoCyc**, (9) KEGG*, (10) iAF1260*, (11) iAF1260$_{deg}$, (12) iAF1260$^{**}_{man}$, (13) Flux-coupling network (fully coupled), (14) KEGG, (15) Intersection of KEGG and iAF12690$_{man}$, (16) Intersection of EcoCyc, KEGG and iAF12690$_{man}$, (17) KEGG**, (18) Flux-coupling network (fully and directionally coupled), (19) iAF1260$^{**}_{deg}$, (20) iAF1260**, (21) Flux-coupling network (directionally coupled), (22) iAF1260, (23) iAF1260*.

links (pairs of reactions sharing a metabolite) into account, which are associated with two single distinct genes (see also Text A and Figure A.2).

(*iii*) We also computed signatures for different intersections of all available databases. By doing so, we gradually remove uncertain connections between genes, nomenclature issues and differences in the level of chemical detail captured by the different databases. This increases the confidence of the used gene network. The intersection of the gene networks from KEGG, EcoCyc and the *i*AF1260 model constitutes hereby the network with the highest confidence as it includes only connections being present in all databases. It should be noted that the differences in the results under variation of the database are also due to the balance between enhancing the systematic contribution (e.g., by eliminating currency metabolites) and retaining a large enough network to extract statistically meaningful quantities.

(*iv*) Different treatments of currency metabolites in case of the *i*AF1260 network (see Text A and Figure A.3 and A.4): (a) manual curation, (b) threshold heuristic and (c) no treatment.

(*v*) Recently, flux-coupling networks have been intensely studied in terms of their organizing principles and their relation to gene expression data. A flux-coupling gene network coming from [86], which has been obtained from the *i*JR904 *E. coli model* [97], is analyzed here. It is subdivided into three subsets: (a) The total network, and two subsets, i.e. (b) fully and (c) directionally coupled gene pairs.

The overall trend seen in Figure 3.5 is that metabolic coherence is highest in the wild type. The mutants' expression patterns, while displaying a positive MC, are not as well aligned to the metabolic network as the wild type. This effect is particularly clear when only switched-on fluxes are taken into account. In this case the metabolic coherence directly measures the coherence of the expression pattern with the pattern of metabolic fluxes. Furthermore, we find a similar pattern for the fully-coupled flux-coupling gene network, which indicates that besides the topological matching also other metabolic relationships are perturbed in the mutants.

Qualitatively speaking, considering intersections and restricting the analysis to fluxes, which are predicted active by FBA, enhances the dominant signal of high wild type metabolic coherence compared to the mutants.

Growth medium complexity

In order to assess the robustness of the result obtained from the flux-activity network shown in Figure 3.4D, it is instructive to analyze how the metabolic coherence (and in particular the strong differences between wild type and mutants) depend on the growth medium: for Figure 3.6 we start out with a rich medium and iteratively remove components until we reach a minimal growth medium. Thus the starting points of the four MC curves in Figure 3.6 coincide with the MC values shown in Figure 3.4D. When going from a rich to a minimal medium, the number of active

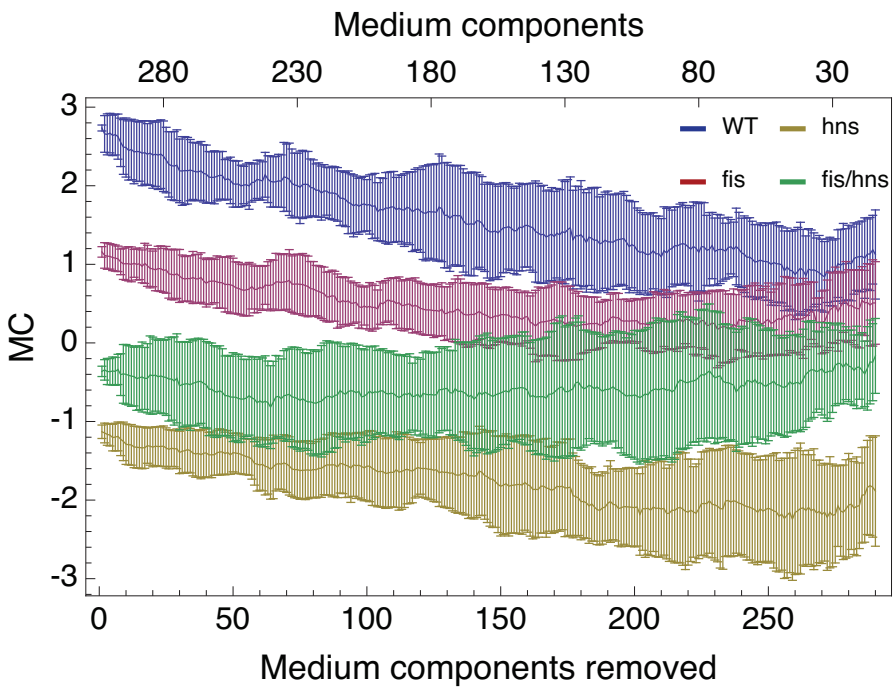


Figure 3.6: MC under varying media conditions. Starting from a rich medium, medium components are removed one by one under the condition that biomass production is not disrupted until a minimal medium composition is reached. Mean MC values over 20 simulations are shown for the wild type (blue), *fis* (red), *hns* (yellow), and *fis*/*hns* double mutant (green) effective gene network. Error bars represent the standard deviation.

genes increases (see Figure A.1), as more and more reactions have to be switched on to compensate for the decreasing nutrient availability. Additionally, from left to right we are deviating ever more strongly from the experimental conditions behind the gene expression data. The main result in Figure 3.6 is that the clear separation of the wild type metabolic coherence from the mutants' persists over a wide range in medium complexity. Furthermore, when approaching a minimal growth medium, discrimination of MCs is strongly reduced.

Link to digital control

Is the strong metabolic coherence found for wild type *E. coli* a direct consequence of chromatin organization (analog control) or is it mediated indirectly through the transcriptional regulatory network (TRN)? From [81] we know that digital control (i.e. the consistency of the analyzed gene expression patterns with the TRN) is low in the wild type (compared to the FIS and H-NS mutants) on the network-wide scale. Here we measure this consistency for a part of the TRN that only consists of reg-

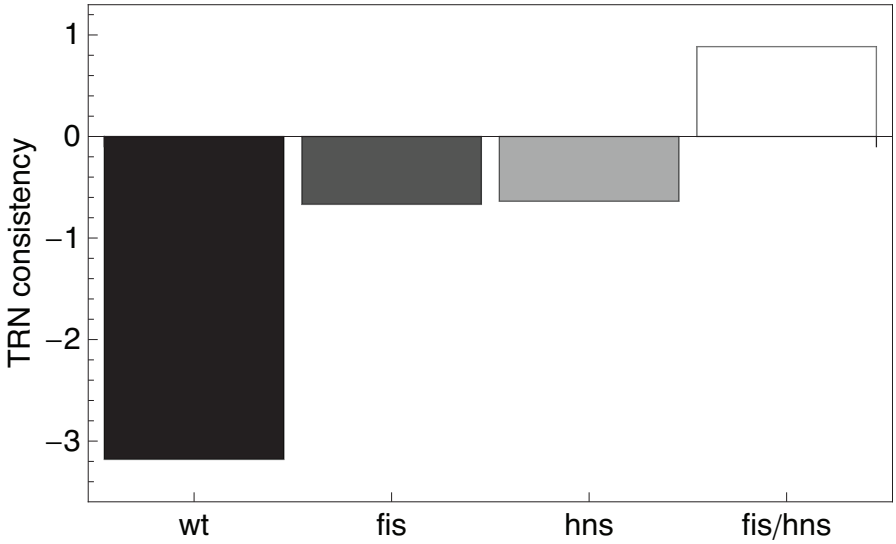


Figure 3.7: Consistency of the signs of supercoiling-induced gene expression changes with the transcriptional regulatory network

ulatory actions (links) between metabolic genes found in the EcoCyc network and genes coding for transcription factors. As expected, the digital control measured as the *digital CTC* ([81]; see also see Text A) is significantly lower in the wild type (see Figure A.5).

Beyond the standard digital control strength from [81] we also integrate the signs of the expression changes with the regulatory information on the corresponding links in the TRN (see *Materials and Methods*, Text A and Figure A.6). This is an elegant method for strengthening the direct link between supercoiling and the metabolic network: not only is the pattern of supercoiling-induced gene expression changes meaningfully distributed on the metabolic network, but also does the transcriptional regulatory network not provide an adequate interpretation of the data (see Figure 3.7).

Conclusions

Our main result, the high metabolic coherence of supercoiling-induced gene expression changes in wild type *E. coli*, as opposed to mutants lacking the NAPs FIS and H-NS, provides strong evidence for a regulatory role of DNA supercoiling. It is robust across several metabolic databases and over a wide range of environmental conditions, when taking flux-activity predictions into account. Furthermore, it is not qualitatively affected by technical details of defining the metabolic network. We can only bring these MC values down by mutations perturbing the machinery of chromosomal organization. These mutants are still viable, but their pattern of supercoiling-

induced gene expression changes shows a markedly reduced metabolic coherence. They are, in fact, close to random expression changes, indicating that changes in the superhelical density cannot be utilized efficiently by the mutants. Furthermore, the low consistency of the wild type expression patterns with the TRN topology (digital control) and its encoded regulatory logic (TRN consistency), suggest that the transcriptional regulation of enzymatic genes is indeed directly enforced by DNA supercoiling.

The results presented here, while providing a fairly clear picture of the interplay between mechanisms of gene regulation and metabolism, provide several incentives for our analysis as obvious steps for future work: at the core of our analysis is the metabolic coherence. It would be helpful to compare this measure with related attempts of quantitatively comparing gene expression data with metabolic information [18]. Also, if suitable data are available, we would like to extend our analysis to other organisms. A more careful discussion of the gene-reaction mapping from a network perspective is certainly necessary in order to go from our observation of metabolic coherence to a more detailed interpretation. It also may be helpful to manually construct metabolite, reaction and gene mappings between *i*AF1260, KEGG and EcoCyc, in order to better understand the strong differences in MC between the databases.

On a broader level, we believe that the general approach of defining and comparing control strengths and topological coherence measures associated with distinct biological processes and, in this way, dissecting gene expression patterns, may be a useful perspective for systems biology investigation, where a multitude of influences shape a process at hand. In those cases where the control type under investigation is network-based (like the metabolic coherence defined here), control strength evaluates effective networks (defined as the currently active part of the static background network). Such effective networks are a novel and highly instructive way of exploring the relation between network architecture and dynamical processes (see, e.g. [76], for an analysis of effective gene regulatory networks and [60], for a theoretical study of effective networks).

Methods

A detailed description of materials and methods is given in Text A.

Gene-centric metabolic networks

We represented metabolism in form of a connectivity network of metabolic genes. We define metabolic genes G as DNA units that encode enzymes or parts of enzyme complexes. Let the gene product of gene $G1$ be involved in reaction $R1$ and that of of $G2$ in $R2$. In the gene-centric metabolic network we study here, the two genes $G1$

and $G2$ are directionally connected if and only if the same metabolite exists among the products of $R1$ and the substrates of $R2$. Networks representing the full metabolism of *E. coli* K12 MG1255 have been constructed from the following sources: the EcoCyc [70] pathways were extracted from the pathways.dat file contained in the flat-file distribution (EcoCyc version 13.6). Neither signaling nor superpathways have been considered in our analysis. The KEGG pathways [69] were retrieved from a distribution of xml files (`ftp://ftp.genome.jp:21/pub/kegg/xml/organisms/eco/`; extracted on 20 November, 2009) describing the different pathways included in the KEGG database. The *in silico* reconstruction *i*AF1260 [39] was obtained in SBML format [58] from the BIGG database [105]. In order to avoid irrelevant connections coming about due to highly abundant compounds, e.g. ATP or other cofactors, sometimes termed currency metabolites [77], we utilized data sources (EcoCyc, KEGG) where these metabolites already have been removed on a reaction to reaction basis. In lack of this information (like in *i*AF1260) we employed a threshold on the metabolites' connectivity degrees to exclude those factors prior to network construction or removed them manually (see also Text A and Figure A.4).

Metabolic coherence

For each effective subnetwork G the ratio of connected nodes to overall nodes was calculated as the metabolic coherence ratio MCR. To make this measure robust against sample size effects we transformed it into a z-score, the metabolic coherence MC, by mapping random gene sets of the same size (i.e. the number of genes/nodes in the effective subnetwork G) onto the overall static network, thus constructing random effective networks G' with associated MCR' values. The MC was computed using 5000 realizations of the null model. A jackknife test was sometimes used to verify the robustness of the MC. The MC was recalculated 100 times while randomly removing 10% of the expression data.

TRN consistency

The *E. coli* transcriptional regulatory network was obtained from RegulonDB ([41]; version 6.4). The consistency of effective TRN subnetworks was calculated as the ratio of consistent links (i.e. the regulatory logic encoded on the links is consistent with the expression signs on the nodes) to overall effective links. Similar to the MC, this ratio was transformed into a z-score. Shuffling the expression signs of the effective nodes was used as a suitable null model. 5000 realizations of the null model were used for the z-score transformation.

Constraint-based modeling

Constraint-based models [95] and especially flux balance analysis (FBA; [128]) and its variants allow the prediction of steady-state flux distributions for genome-scale metabolic models by solving a linear optimization problem under various subsidiary conditions. This approach has been used thoroughly in the past to tackle a wealth of questions regarding the metabolic capabilities of different organisms [71, 95]. For the computation of flux distributions under varying media conditions, we started from a rich medium, removing medium components one by one under the condition that biomass production is not disrupted until a minimal medium composition was reached. We should mention that a large number of trajectories through the traversed media space exists.

Experimental setup

The transcript profiles analyzed in this study were obtained by DNA microarray analyses using genetically engineered *E. coli* LZ41 and LZ54 strains containing norfloxacin-resistant topoisomerase gene alleles to selectively inhibit either DNA gyrase or topoisomerase IV activity and respectively induce either relaxation or high negative supercoiling [140]. Introduction of the *fis* and *hns* mutations in the LZ41 and LZ54 strains did not alter the global supercoiling response to drug addition was not substantially [20]. By adding norfloxacin to the LZ41 and LZ54 strains and their mutant derivatives we could vary the superhelical density σ in opposite directions and distinguish gene transcripts associated either with relaxation ($\downarrow \sigma$) or high negative supercoiling ($\uparrow \sigma$) in each genetic background. ArrayExpress accession numbers: E-MEXP-462 and E-MEXP-463.

Authors contributions

NS, MG, GM and MTH designed research. NS, MG performed research. NS, MG analyzed data. NS, GM and MTH wrote the paper.

Acknowledgements

We would like to thank Balázs Papp for helpful discussions regarding the flux-coupling analysis. Furthermore, we want thank Miriam Grace and Mortitz Beber for helpful comments on the manuscript.

Chapter 4

Multiple topological labels and medium-dependent essentiality in *Escherichia coli* metabolism

This chapter provides the content of the following manuscript in preparation [118]:

> Nikolaus Sonnenschein, Carsten Marr, and Marc-Thorsten Hütt *Multiple topological labels and medium-dependent essentiality in Escherichia coli metabolism*

Abstract

Background: Metabolism has frequently been analyzed from a network perspective. A major question is, how network properties correlate with biological features like enzyme essentiality. We investigate the metabolic system of *Escherichia coli* on the basis of reaction categories that were discovered by different topological and flux balance modeling methods in the past, namely reactions associated with uniquely produced and uniquely consumed metabolites, reactions changing the synthetic accessibility of the biomass and metabolic core reactions. Individually considered these reaction categories all reveal a high amount of reactions essential for the production of biomass.

Results: We performed a large-scale flux-balance analysis simulation on random media to explore two lines of study based on topologically motivated reaction categories: (1) medium-dependent essentiality may clarify the distribution of essential reactions into the different categories and provide insight into the biological significance of the topological classifications; (2) using logical operations on the intersecting sets of reactions may improve topology-based essentiality prediction. In particular, we observe that medium-dependent essentiality is more highly correlated with synthetic accessibility than with the other reaction categories.

Conclusions: Our method of using multiple topological labels for investigating the relation between network properties and system properties and in this way identifying sub-categories emanating from different topological features may be helpful in a broad range of contexts in systems biology. In the case of metabolism, we observe that some combinations of reaction categories show a substantially higher accuracy in predicting essentiality, suggesting different types of essential reactions.

Background

The question, how network topology shapes dynamic processes, is currently under intense investigation in a wide range of disciplines – from technical [48, 137] and social [4, 49, 89] systems to biology [44, 96]. With the advent of "network biology" [17] in the late 1990s and early 2000s [16, 133], metabolism was among the very first intracellular networks studied from a topological perspective [5, 47, 50, 64, 131].

Along this line of research, metabolic reactions have been classified in several different ways based on topological information [5, 52, 79, 131]. Here we will focus on two recent examples providing such classifications: UP-UC reactions and SA reactions.

UP-UC metabolites have been introduced in [104]. UP-UC metabolites were described as metabolites that are consumed and produced by only a single reaction and, thus, exhibit the lowest possible degree in a bipartite network representation of the system. A UP-UC cluster may then be defined as a reaction subset that connects a set of UP-UC metabolites. Besides the high essentiality of this UP-UC reactions, which is the key issue in this work, this metabolic reaction category comprises also some other quite interesting features like proportionally fixed flux values under steady-state conditions or their correlation with regulatory modules [104].

The synthetic accessibility (SA) measure has been defined in [136] and is influenced by a measure used in chemical drug design describing the number of steps needed to synthesize a specific compound from a given set of common laboratory reactants. Accordingly, the SA for a metabolic system is defined as the minimal number of reactions needed to reach a set of outputs (e.g. biomass) from a given set of

inputs (e.g. medium composition) as obtained by a breadth-first-search traversal that can only proceed if all needed substrates are available. SA is successful in predicting essential genes as lethal mutations cause a change in the SA [136]. For this work we choose to treat SA as a reaction category by assigning an SA label to every reaction whose knock-out causes a change in biomass SA.

Figure 4.1 shows a schematic representation of metabolism with three exchange reactions ($X1$, $X2$ and $X3$) with the environment and a two-component biomass reaction (BM). Circles represent metabolites, while boxes stand for enzymes in this bipartite graph view of a metabolic system. In this figure, $E1$ (highlighted in blue) is an example of an SA reaction, as it represents one of the shortest paths to BM, while $E5$ (highlighted in green) is consuming and producing only metabolites, which are uniquely produced (UP) and uniquely consumed (UC), and thus is an example of a UP-UC reaction. Figures 4.1a-c provide a qualitative impression of the wild type flux distribution (Figure 4.1a) and the re-routing of fluxes upon $E1$ and $E5$ knockout (Figures 4.1b,c), respectively.

In the example in Figure 4.1, both reactions ($E1$ and $E5$) have an alternative path that goes along reaction $E4$. Thus, both reaction labels would in this case not serve as a reliable predictor of the reaction's essentiality. Eliminating reaction $E4$ (see Figure 4.1d) from the system would remove this alternative path, thus turning $E1$ and $E5$ into correct essentiality predictors. This (suggestive) example illustrates, why a systematic study of combinatorial subsets of these categories can be interesting for understanding the topological basis of essential reactions.

A third category of reactions, which has recently been introduced, comes from taking into account some information on the flux distributions as well: A sampling of steady states predicted from flux-balance analysis over a wide range of randomly chosen media conditions standard minimal growth medium) revealed a metabolic core (MC), which is always switched on [7, 8].

Remarkably, these different classification schemes or categories are all fairly accurate predictors of essential reactions (i.e. reactions, where a mutation in the gene encoding the corresponding enzyme is experimentally known to be lethal). For the MC this is intuitively clear: The set of reactions always switched on under a wide range of media should contain a large number of essential reactions. Both, in the case of SA and UP-UC reactions, one can argue that, topologically, they constitute "bottleneck paths" without (at least local) detours providing alternatives (cf. also Figure 4.1).

Here we study the overlap of these different reaction categories and the question, whether performance in predicting essentiality can be increased by combining the different categories in a combinatorical fashion. We furthermore introduce a more refined, simulation based measure of essentiality, i.e. the relative essentiality of a reaction.

The methodological framework of our study is flux-balance analysis. Mathematical optimization methods formerly used mainly in engineering, economics, physics

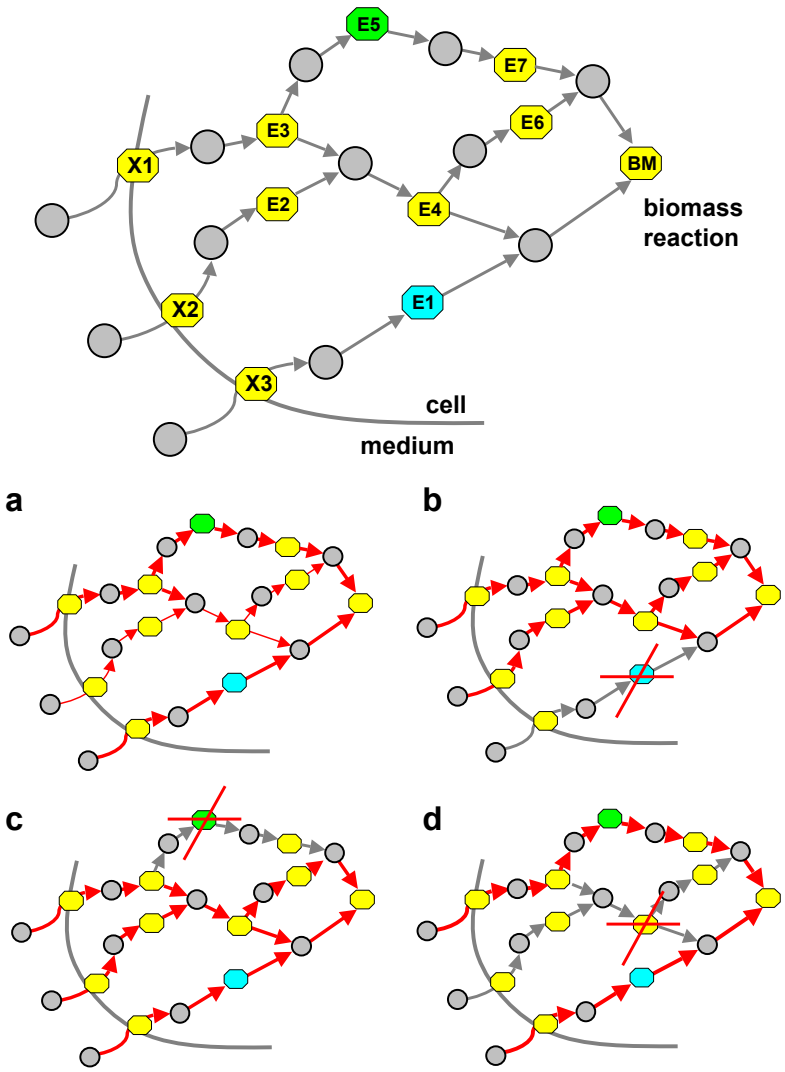


Figure 4.1: UP-UC and SA reactions. Simple scheme of a small fictitious metabolic reaction system with examples of UP-UC and SA reactions are (large picture). (a) Wildtype network. (b) Knockout of SA reaction $E1$. Fluxes are rerouted over $E4$ leading to an increase in the systems SA. (c) Knockout of UP-UC reaction $E5$. (d) Knockout of reaction $E4$. $E1$ (SA) and $E5$ (UP-UC) are now correct essentiality predictors.

and chemistry are becoming increasingly popular in the field of systems biology and the applications range from model building and parameter estimation to optimal experimental design and synthetic biology [15].

Particularly its capacity to predict gene essentiality with high accuracy for *E. coli* and *Saccharomyces cerevisiae* has turned flux-balance analysis (FBA) into a widely accepted method for *in silico* studies of metabolic states [36, 95, 128].

The conceptual starting point of FBA is the stoichiometric matrix S, which relates the vector v of metabolic fluxes with changes dm/dt in the vector m of metabolite concentrations $dm/dt = Sv$. Assuming a steady state, $dm/dt = 0$, one obtains a homogeneous system of linear equations. Furthermore assuming bounds for each of the uptake rates of the compounds from the medium one obtains a well-defined subset of flux space, the "flux cone", as a potential solution space. The optimal solution is now identified by maximising a given objective function, e.g. the biomass production of the organism.

A variety of implementations of FBA are by now available in the systems biology community [19, 126], as well as for research in biochemical engineering and other biotechnological contexts, helping to make FBA a very successful tool for studying this otherwise (on the system-scale level) rather elusive level of cellular organization.

Recent refinements of FBA focus on the distribution of fluxes upon mutations [108, 110], the incorporation of temporal information beyond the steady state [71, 80] and of gene regulatory information [29, 103, 111]. The recent observation, however, that the majority of fluxes determined by metabolism alone is consistent with those obtained by evoking additional gene regulatory input [111] strengthened the case for the use of FBA for large-scale system-wide studies of metabolism.

Although experimental data from systematic knockout studies is available for *E. coli* [12, 43] these essentiality profiles result from a limited set of environmental conditions. In particular, it has been pointed out recently that essentiality is often medium-dependent [54, 93]. While this has been analyzed in [54] for genetic interactions (i.e. the effect of a knockout under the condition of another knockout) we analyze here the above categories (SA, UP-UC and MC reactions) in the light of single-knockout medium-dependent essentiality.

Results and Discussion

Overlap of the three reaction categories

For all three reaction categories discovered here a high overlap of the retrieved subsets with the set of essential reactions was reported, suggesting that also the intersections among these categories should be rather high. To study this assumption we analyze the UP-UC, SA and MC pairwise intersections and compare them with the intersec-

Table 4.1: Overlap of the three reaction categories. The overlap between the UP-UC, SA and MC reaction categories versus the expected overlap based on two different reaction pools, together with the corresponding z-scores.

Reaction Pool	Intersection	Real overlap	Expected overlap	z-score
All classified reactions	UP-UC ∩ SA	0.518	0.244 ± 0.029	9.4
	UP-UC ∩ MC	0.604	0.232 ± 0.040	9.3
	SA ∩ MC	0.719	0.244 ± 0.041	11.6
Only essential reactions	UP-UC ∩ SA	0.727	0.689 ± 0.030	1.3
	UP-UC ∩ MC	0.611	0.534 ± 0.038	2.0
	SA ∩ MC	0.726	0.689 ± 0.035	1.1

tions of randomly drawn sets based on two reaction pools: (i) all reactions, for which a characterization into "essential" and "non-essential" is available (724), (ii) all essential reactions (206). The results are summarized in Table 4.1. For consistency in all cases only the reactions appearing in in these two pools have been considered.

The comparison reveals two features of the overlaps: First, the overlaps are far larger than expected at random based on pool (i), suggesting that the effective pool of reactions is substantially smaller. Second, when restricting this pool to the essential reactions only (ii), the overlaps still deviate systematically (but far less strongly) from the expected towards higher values. This latter point suggests that a sub-classification of the essential reactions might exist, where each sub-class matches more closely each pair of reaction categories and therefore accounts for the remaining enhancement of the overlap. In the following we explore the possibility that relative essentiality provides such a sub-classification.

Relative essentiality: An illustrative example

In the following we use two definitions of essential reactions. The first are the experimentally determined essentiality profiles, as reported in [12, 43]. These data sets have also been used in many previous studies on essentiality prediction [6, 8, 67, 88, 104, 136]. The second definition of essential reactions is based upon the FBA prediction of essentiality, which, in the case of *E. coli* agrees well with the empirical data (92 % accuracy under glucose aerobic growth conditions, see [39]). Due to the enormous combinatorical range of potential media, the important class of medium-dependent essentiality, similarly to the MC [7, 8] (which also requires a large-scale sampling of media space) is only accessible via FBA prediction.

In need of a global quantity we determined the essentiality profiles computationally for 6×10^5 random media conditions and yielded a continuous measure of relative

essentiality for every reaction in the system. Essential reactions were determined by an assumption of having a growth rate at most 5% of the wild type [93]. We define the relative essentiality as the percentage of cases, in which the removal of a specific reaction was lethal out of all simulated environments, where the reaction was active.

As an illustrative example of medium-dependent essentiality (i.e. relative essentiality arising from monitoring essentiality across a large set of media) we want briefly discuss the central metabolism model for *E. coli* (*E. coli* core model, see [90]) as a minimal system and study the predicted biomass production under knockout for seven different carbon sources (glucose, acetate, fumarate, lactose, succinate, ethanol, pyruvate). The FBA model consists of 63 reactions, including 14 exchange reactions, which also regulate the uptake of the respective carbon source provided. FBA studies for this system have been performed using the 15 component biomass function provided with the model [90]. Figure 4.2a shows the size of the biomass flux for different mutants (bars) for each medium (color segments in each bar). As expected, the biomass flux is typically largest for the glucose medium. In one case (i.e. the removal of *GLCpts*, which is a glucose transport reaction), however, only other carbon sources lead to a non-zero biomass flux prediction. The relative sizes of the color segments (i.e. of the biomass flux under a particular carbon source) vary greatly from mutant to mutant.

For example the removal of the reaction FBP, which is catalyzed by the enzyme fructose-1,6-bisphosphatase and dephosphorylates fructose-1,6-bisphosphate to fructose-6-phosphate is essential for all carbon sources except glucose. As a matter of fact, FBP is a step in gluconeogenesis, which is a pathway that generates glucose. In the case of the glucose medium there is no need to generate glucose from other substances as it is provided directly via the glucose uptake reaction.

We have to mention that *in vivo* the deletion of the gene fbp is not lethal in *E. coli* as it can bypass the loss through other pathways. The deviant result comes about due to the restricted minimal model used in this illustrative example covering only central metabolism. As we proceed to the genome-scale reconstruction *i*JR904 we find that FBP is actually never essential, in agreement with experimental results [12, 43]

Relative essentiality analysis for the full system

In order to subdivide the metabolic reactions into essentiality classes, namely essential (persistent-lethal), non-essential and partially essential (or conditional-lethal), we quantify a reaction's *relative essentiality* by simulating 6×10^5 random media conditions and performing all single reaction knockouts (leading to $> 1.8 \times 10^8$ individual FBA calculations) to identify for each medium the set of essential reactions.

The relative essentiality is then defined as the number of lethal outcomes upon the removal of a target reaction divided by the number of environmental conditions this reaction was active. An alternative definition of relative essentiality would be to

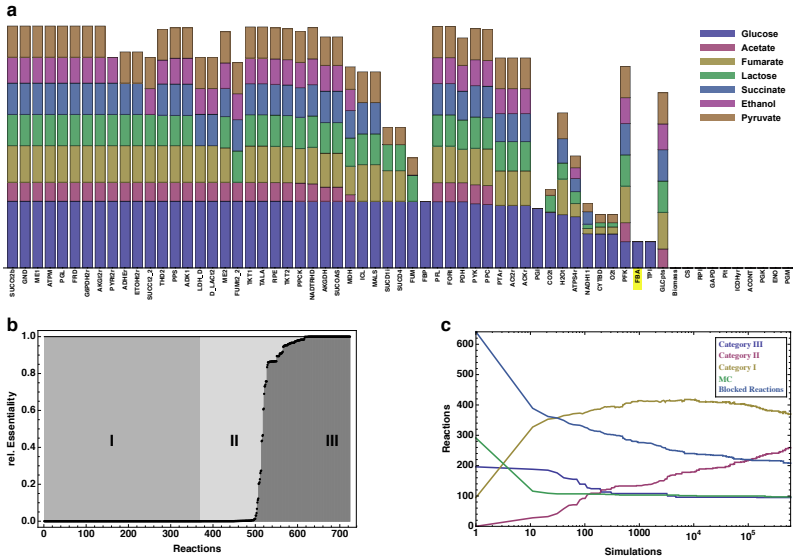


Figure 4.2: Medium-dependent essentiality. (a) *E. coli* core model - medium dependent essentiality for seven carbon sources under aerobic conditions. The height of the bars indicates the size of the growth flux as determined by FBA. The reactions have been sorted according to (b) Sorted global essentiality profile determined by the simulation of 6×10^5 random media conditions. The three different essentiality classes are indicated by different background colors. Class I (under no circumstance lethal) reactions are indicated in dark gray. Class II (conditional-lethal) and III (global essential) are colored in gray and light gray respectively. (c) Simulation progress of the numbers of reactions in each class. Two additional sets, the MC (i.e. the set of reactions that is always active), and the set of reactions that have not become active yet, are also depicted in the diagram.

normalize the number of lethal outcomes to the total number of media sampled. In this case, however, reactions only active in very few, specialized media would give a very low essentiality value, making it difficult to assess rare reactions.

Figure 4.2b shows the sorted relative essentialities for the available 724 reactions (comprising all reactions in the *E. coli* model, which have been active at least once in all FBA simulations; blocked reactions [24] have thus been eliminated; see also Methods). In Figure 4.2a the three essentiality classes are clearly visible: The removal of most reactions has no or only small consequences for the production of biomass (Class I). Some reactions are globally essential (Class III) and a third set is only conditional-lethal (Class II). Figure 4.2c depicts the change of the category sizes over simulation number. Additionally the MC size evolution and the decrease of the blocked reaction set are included in this figure. The MC and Class III reactions converge pretty fast to the sizes of 96 and 94, respectively. Two reactions from the core are not globally essential due to the fact that other reactions can replace them and rescue

growth although not as high as the optimum the wild type achieves [8]. The slope of the unused reaction set curve becomes rather small after a long simulation time. Eventually, it should reach the size of the computationally determined set of reactions that cannot bear a flux under no circumstances (blocked reactions; see Methods). We decided to exclude reactions in this set from our analysis.

From Figure 4.2c it is evident that even after this large number of random media the size of the Class II and I sets have not yet converged (in contrast to Class III reactions). From approx. 10^4 media onwards, the increase in Class II reactions corresponds to the decrease in blocked (or unused) and Class I reactions, i.e. reactions, whose removal was lethal only under an extremely small number of environmental conditions. As these cases probe rather extreme features of the FBA model, it might be suitable to exclude them from consideration. Similarly, one could argue that slight deviations from zero and one in the relative essentiality may be allowed for Class I and Class III reactions. We decided, however, not to include such thresholds here.

Using this global essentiality measure we determined the amount of reactions belonging to the three essentiality classes for every of the three reaction categories. The results in Figure 4.3 show that the three categories incorporate different amounts of reactions belonging to each of the three essentiality classes. The SA reaction set seems to be composed of a mixture of Class II and III (conditional and persistent-lethal reactions) whereas the UP-UC reactions exhibit a surprisingly high amount of Class I reactions (never lethal). As expected from the definition, which requires the reactions to be always active, the MC category is almost exclusively made up of Class III reactions (94 of 96 reactions belong to Class III).

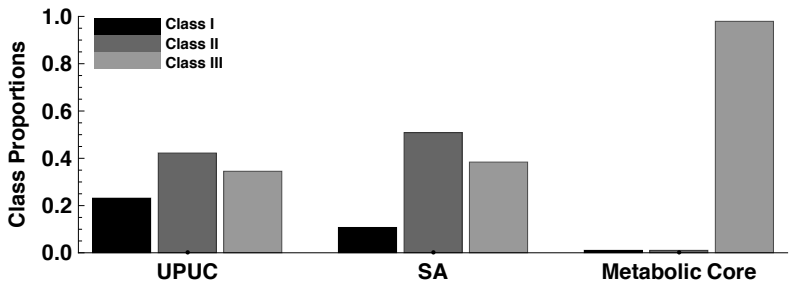


Figure 4.3: Reaction categories and essentiality classes. The proportions of the three different essentiality classes determined for UP-UC, SA and MC component. Class I is indicated in black, Class II in dark gray and Class III in gray.

Reaction categories as predictors of essentiality

We analyzed the performance of the three reaction categories to predict the essentiality profiles of experimental data [12, 43] and the relative essentiality we obtained

computationally by sampling a large set of random media. For a first assessment of the performance and comparison with previous results from [8, 104, 136], we set an arbitrary threshold of 0.5 relative essentiality to map the conditional-lethal reactions of our computational analysis to either Class I or Class III, resulting in the values shown in Figure 4.4a. For the definition of the performance indices, see Methods.

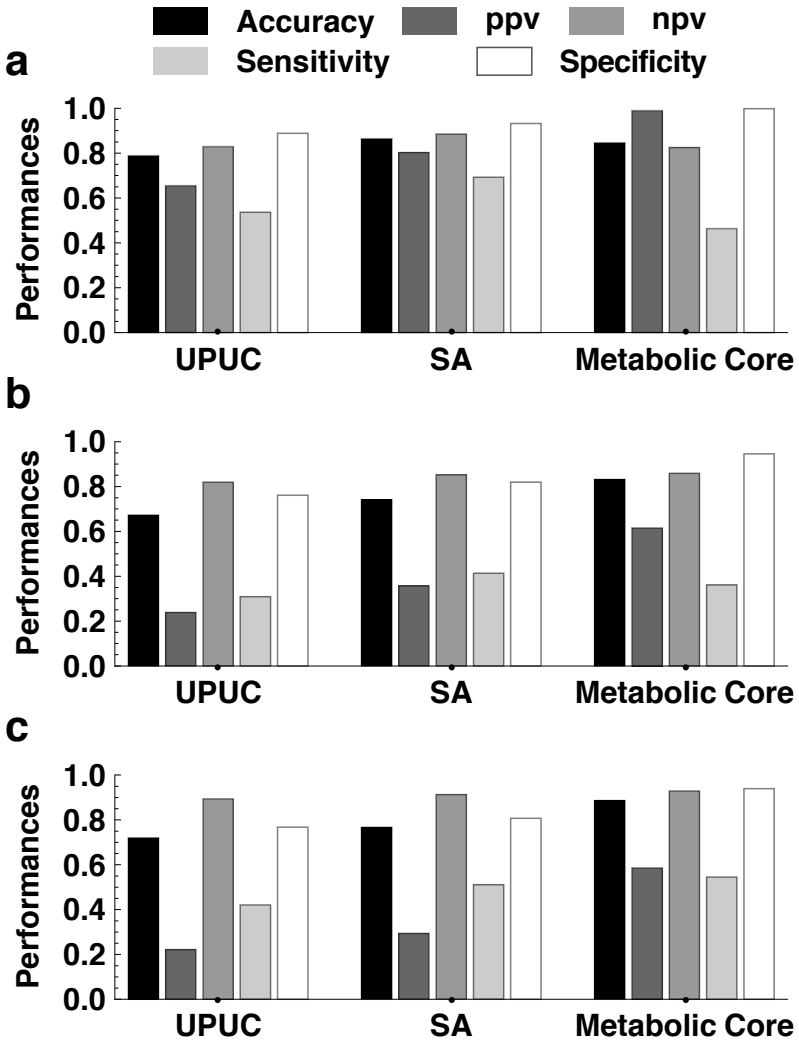


Figure 4.4: Prediction performance measurements. Prediction performance of the three reaction categories on (a) our computational essentiality profile, (b) the Gerdes data set and (c) the Keio collection.

The result for the performance analysis for the two experimental data sets [12, 43] is shown in Figure 4.4b,c. In case of the experimental profile, it is remarkable that the high accuracy values come about almost exclusively through the high non-lethal

prediction rate, as seen in the high negative predictive value and specificity. Both, the positive predictive value and the sensitivity, are rather low although the comparison of the two experimental profile reveals higher sensitivities for the Keio collection [12].

Comparing the performance profiles of the experimentally observed essentialities (Figure 4.4b,c) with the computational predictions (Figure 4.4a) it is particularly striking that the positive predictive value (ppv), as well as the sensitivity, are higher in the computational case. We checked that this effect is not induced by the mere difference in numbers of essential reactions (166 for the Gerdes set, 112 for the Keio set, vs. 206 for the computational set at a threshold of 0.5 in the relative essentiality, cf. Figure 4.2b). If we exclude most conditional-lethal reactions from the computational set (by increasing the threshold to 0.95, leading to 157 essential reactions) one still finds a strongly elevated ppv compared to the experimental sets.

Although all reaction categories display a rather high accuracy (Figure 4.44), the reasons may be individually slightly different: The MC category for example achieves high accuracy dominantly by its high positive predictive value, while the other two categories tend to achieve this by a high sensitivity (particularly in the computational set).

In order to further rule out side effects from the different size distributions of the essentiality data sets we checked our results with a suitable null model. According to the sizes of the reaction categories (UP-UC, SA and MC) we assigned repeatedly the essentiality label to random reaction samples out of all available reactions and computed the accuracy. The accuracy values for the computational set and the Gerdes set are shown in Figure 4.5a, together with the corresponding null model results. For the computational set, the real accuracy is significantly higher than the null model value for all three reaction categories, while for the Gerdes set, the UP-UC category fails to exceed the null model prediction (see Methods for details on the null model accuracy). It is clear that this observation depends on the choice of the null model. Other null models could still be exceeded by the real accuracy values (the null model from [136], e.g., is normalized to always yield an accuracy of 0.5).

In Figure 4.5b we plot the different real and random components in a positive and negative predictive value plane, similarly to [136], to make our previous results more transparent. With the exception of the UP-UC component and the experimental essentiality profile all the other data sets are good predictors of essentiality in comparison to the random model. As in Figure 4.5a, the plane from Figure 4.5b includes only the computational set and the Gerdes set.

Multiple labels

Next, we analyze whether it is possible to increase the accuracy of the essentiality prediction by combining the reaction categories.

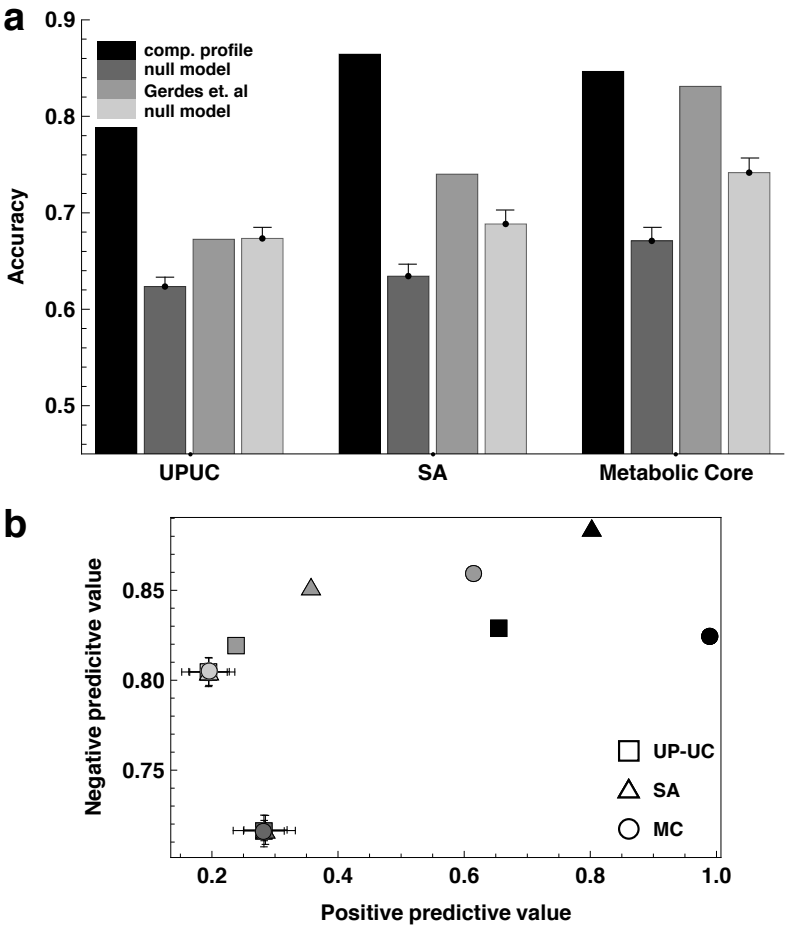


Figure 4.5: Evaluation of the performance measurements. Comparison of the reaction components' accuracies and their random model counterparts. (b) The different real and random reaction components' positive and negative predictive values plotted in a phase plane. The variation in the random models predictive powers are indicated via error bars representing the standard deviation of the simulation.

We determine for all 127 possible combinatorial intersections the values of all performance indices. The ten highest scorers with respect to the accuracy measure are summarized in Figure 4.6a. The highest accuracy was observed in a set containing 205 reactions including almost all reactions of all three components with the exception of the UP-UC-only reaction set (MC ∪ SA), as the latter includes the highest amount of false positives (Figure 4.6a). This is an improvement of 12 %,4 % and 6 %, respectively, in comparison with each single reaction category (UP-UC, SA and MC).

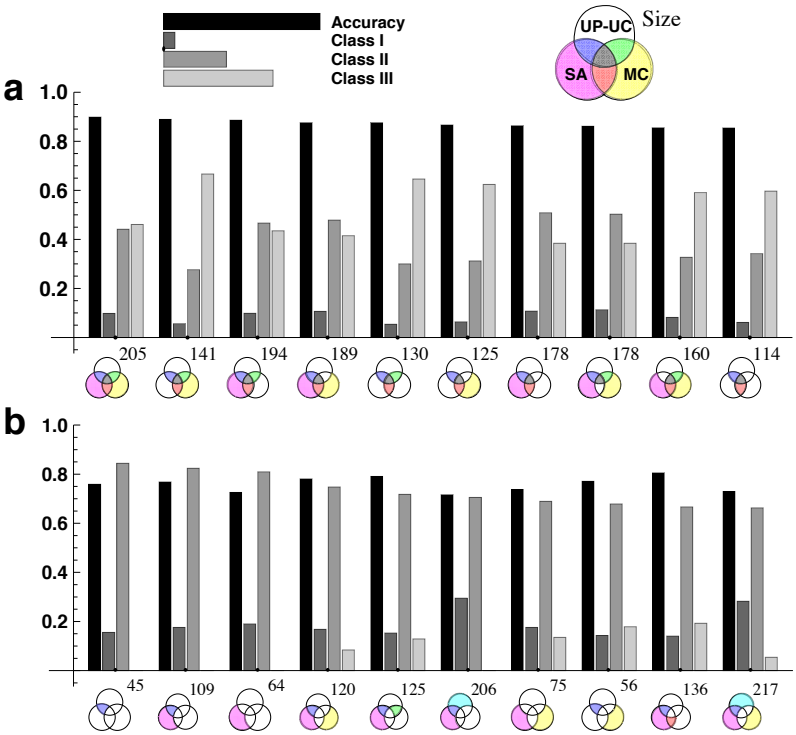


Figure 4.6: Multiple labels - Top 10 performance scorers. (a) The 10 highest accuracy scorers. The black bar represents each the accuracy, ppv, npv, sensitivity and specificity of the intersection configurations. The other colors represent the three essentiality classes with Class I being colored in dark gray and Class II and III in gray and light gray, respectively. (b) The top 10 category II scorers. The black bar represents always the accuracy. The other three bar types represent the three essentiality classes with Class I being colored in dark gray and Class II and III in gray and light gray respectively.

The highest positive predictive value and specificity was achieved with (MC ∩ SA) ∪ (MC ∩ UP-UC) consisting of 85 reactions and it is remarkable that all the other high scorers in these categories are only a variation of this pattern (additional file 1: Supplementary Figures B.1a,d). Furthermore SA-only and UP-UC-only reactions are never a part of them. This is what we anticipated, as the positive predictive value captures the true positive ratio, and in order to reach a high value, large numbers of

false positives have to be avoided. As false positives have to be avoided also for a high specificity the patterns of intersections resemble each other. The highest negative predictive value and sensitivity was achieved with a set consisting of the union of all reaction components, 302 in numbers (additional file 1: Supplementary Figures B.1b,c). At least for the sensitivity this is a trivial result as the theoretical expectation value increases with the number of positive predictions.

As a last step, we want to find out, whether this multiple reaction labeling (and the subsequent analysis of all combinatorially possible sets) reveals clearly characterized subclasses of essential reactions. The analysis of the overlap between the three reaction categories (Table 4.1, cf. the discussion at the beginning of this section), together with the observation that these categories are good predictors of essentiality (Figure 4.4- 4.5; see also [8, 104, 136]), suggests that such subclasses may exist. It is therefore particularly interesting to extend the multiple-label analysis to the three essentiality classes introduced above.

For each combinatorial combination of the reaction categories we determine the intersection configurations with the highest relative proportions of the three essentiality classes. In Figure 4.6b the ten highest scorers are depicted, with respect to Class II content. It is not surprising that the intersection configuration with the highest proportion of Class I reactions (additional file 2: Supplementary Figure B.2a) is exactly the UP-UC-only set (i.e. UP-UC $\cap$ SA$'$ $\cap$ MC$'$), because we already have shown in the previous paragraph that exactly this set is excluded from the intersection configuration with the highest possible accuracy. Similarly (and not surprisingly), for Class III reactions (additional file 2: Supplementary Figure B.2b) the MC category dominates the high scorers; this is in agreement with the findings of the previous paragraph and also results in Figure 4.3, as the MC component contains the highest amount of Class III reactions.

For Class II reactions, we find it quite interesting that the highest proportions in this class are seen in an UP-UC–SA only intersection (i.e. UP-UC $\cap$ SA $\cap$ MC$'$) (Figure 4.6b) and all the other high scorers resemble this pattern with the SA component playing the major role for the outcome of such high Class II proportions.

A typical Class II reaction (approximately 20 percent of all Class II reactions are of this type) is SA but not UP-UC and not MC (i.e. SA $\cap$ UP-UC$'$ $\cap$ MC'). Among the (SA $\cup$ UP-UC) $\cap$ MC$'$ reactions we find approximately 15 percent of all Class II. In contrast, Class III reactions are spread evenly among SA and UP-UC.

A reaction is classified as SA, when the biomass is more difficult to reach after the knockout of that reaction, even if (longer) alternative paths to the biomass exist. The availability of such alternative paths depends often on the specific medium provided. This common assessment of alternative paths could explain the observed high predictive power of the synthetic accessibility for Class II reactions.

In spite of its topological basis, the synthetic accessibility combines local and global graph features and in this way gets, seemingly, closest to matching medium modifi-

cations, which also have local and global features by addressing typically whole paths through the system. Qualitatively speaking, a large percentage of Class III reactions (always essential) may stand out topologically, i.e. they may be identifiable from local graph properties alone. Due to the interplay of local and global properties, we expect this to be less valid for Class II reactions. The UP-UC category can thus be expected to work better for Class III reactions than for Class II.

Conclusions

By analyzing the range of predictions for essentiality from three established topological (or topology-motivated) labels we attempted to better characterize the specific topological constellation that makes a reaction essential.

The concept of medium-dependent essentiality (see also [88, 93]) is particularly instructive in this context of multiple labels: This intermediate regime between essential and unessential reactions, which is not very visible in any of the three labels specifically, stands out significantly in particular label combinations.

We suspect that the reason for the significant over-representation of conditional-lethal reactions in the SA component lies within the intrinsic features of the synthetic accessibility concept itself, efficiently combining local and global features of the metabolic reaction network.

It might be that the essentiality predictions carried out by the synthetic accessibility approach [136] depend on the set of inputs and bootstrapping metabolites that are chosen. It would be interesting to explore this point in more detail.

We have not included the results from [74] in our analysis, which aim at metabolite essentiality. The high essentiality of UP-UC metabolites can be in principle recovered from their results and thus a comparison may be worthwhile, even though [74] would yield a metabolite category, rather than a reaction category.

It would be particularly interesting to extend the range of reaction categories, in order to characterize typical essential reactions even better from a topological perspective. Naturally, other purely topological properties (based on the degree of a node or the betweenness centrality) may be included, but also, e.g. the metabolite scopes described in [52] could be in principle an extension to our labeling approach.

Multiple labeling could also be a helpful tool in the combination of graph theoretical and high-throughput data. The recent study [86] on flux coupling as a predictor of correlations in gene expression compared to a different (but often related) observable, the distance of reactions, could be an interesting field of application of such a multiple labeling approach, in order to specifically see, how the high-distance but flux-coupled reaction pairs (and other subsets within the two categories) perform in terms of correlation to gene expression.

Methods

Model

The genome-scale metabolic reconstruction *i*JR904 [97] of *E. coli* was used in all our experiments. Each reversible reaction was replaced by two irreversible reactions acting in opposite directions. For topological analyses a bipartite graph representation was generated from the stoichiometry. Such a graph consists of metabolite and reaction nodes connected by directed links, where substrates point towards reactions they are consumed by. These reaction nodes then again point towards their products. No direct connections exist within the reaction and metabolite subsets.

Flux Balance Analysis

Linear programming (LP), as the algorithmic backbone of flux balance analysis (FBA) [128] and its variants make the computation of steady state flux distributions of genome scale models possible and has been used in the past to tackle a wealth of questions regarding the metabolic capabilities of different organisms. [71, 95]

Generally, the LP problems defined in FBA can be stated as follows:

$$\begin{array}{ll} \text{Maximize} & Z \\ \text{subject to} & Sv = 0, \\ & v_{min} <= v <= v_{max} \ . \end{array} \tag{4.1}$$

with an objective function Z, the stoichiometric matrix S, the flux vector v and the constraint vectors v_{min} and v_{max}. As we are considering reversible reactions as two independent unidirectional reactions, we set v_{min} to zero. This LP problem can be solved through linear optimization and, as the solution space is convex, a global maximum can be computed, if it exist, although multiple optima cannot be excluded. All LP problems throughout our analyses have been solved using custom Python bindings for the C API of the GNU Linear Programming Kit (GLPK).

Blocked reactions

We removed all globally blocked reactions from the model to give the topological methods described in this article (UP-UC, SA) the opportunity to work on the same information content as their dynamical counterpart (MC). A high (not as high as the default flux boundaries v_{max}) maximal uptake and secretion rate was assigned to all available transporters in the system and then blocked reactions were confirmed by flux variability analysis [24]. These globally blocked reactions cannot carry a flux under any environmental conditions and consequently are not available to methods that use FBA.

Metabolic core reactions

The MC reactions were computed similarly to the procedure described in [8]. Random media compositions were generated by picking randomly between 10 to 100 percent of all available transport reactions in the model and the assignment of random uptake and secretion rates in the interval of 0 to 20 $mmol/gDWh$. For all of these random configurations the biomass objective was maximized, and in the case of a resulting growth flux below 0.5 $mmol/gDWh$ the result was discarded.

Synthetic accessibility reactions

The synthetic accessibility of all reactions in the system was computed according to [136]. The needed outputs were defined to be the substrates of the biomass function and the ingredients of a glucose minimal medium were defined to be the inputs of the system.

As a variation to [136] we decided to include no further additional compounds that ensure that all outputs are reached in the wild type. Instead we used a set of bootstrapping metabolites [100] that permit a proper functioning of the algorithm but are not the starting points of the breadth first search.

UP-UC reactions

The UP-UC reactions were determined in analogy to the algorithm published in [104]. We determined all metabolites with an in-degree and out-degree of one (UP-UC metabolites) in the bipartite graph representation of the metabolism of *i*JR904. Then we computed the set of reactions (UP-UC reactions) that are associated with the set of UP-UC metabolites for further analysis.

Experimental and computational essentiality profiles

We used the essentiality profiles published in [12, 43] for our performance analyses. The computational essentiality profile was determined by extending the MC analysis with single reaction deletions. For every flux distribution along the random media sampling procedure we determined the set of active reactions and checked their essentiality by knocking them out one by one. A knockout was achieved by constraining the flux of the specific reaction to zero and repeating the optimization under the same random environmental conditions. To transform our continuous essentiality measure into a binary form, whereas true represents lethal and false viable we set a threshold of 0.5 relative essentiality to map the reactions in Class II (conditional-lethal) to either Class I or II.

We find 200 UP-UC reactions (for 168 of which have essentiality information), 178 SA reactions (177 with essentiality information), and 96 MC reactions (96 with essentiality information). The numbers provided in Figure 4.6 refer to the full sizes of reaction categories. Classes I, II and III consist of 370, 259 and 94 reactions, respectively.

Assessment of predictive power

We have quantified the predictive capabilities of the different subsets by five observables from statistics relying on binary classification, namely the accuracy $(TP + TN)/(TP+TN+FP+FN)$, positive predictive value $(TP)/(TP+FP)$, negative predictive value $(TN)/(TN + FN)$, sensitivity $(TP)/(TP + FN)$ and specificity $(TN)/(TN+FP)$. In all these observables TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. If we denote the essential reactions in the real data sets by + (and inessential reactions by −), as well as the reactions belonging to the category under consideration (and therefore serving as predictors of essentiality) by p (with n denoting all reactions not in this category), we can e.g. represent TP as the joint probability $P(+, p)$ of + and p. In our null model we then have $TP = P(+)P(p)R$, where $P(+)$ is the percentage of essential reactions, $P(p)$ is the relative size of the reaction category and R is the total number of reactions. Consequently, our null model accuracy is given by $P(+)P(p) + P(-)P(n) = (P(+) - P(-))S_R/R + P(-)$ with the number of reactions S_R in the reaction category. Thus, our null model accuracy depends on the size of the reaction category, unless if about half of the reactions are essential.

Authors contributions

NS and MH conceived the study and wrote the manuscript. NS carried out the analyses. CM participated in the design of the study and contributed to the statistical analysis. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Zeba Wunderlich for providing additional information on the synthetic accessibility approach. Furthermore the authors like to thank Areejit Samal for discussions about the topological implications on the essentiality of reactions.

Chapter 5

A network perspective on metabolic inconsistency

This chapter provides the content of the following manuscript in preparation [116]:

> Nikolaus Sonnenschein, Arndt Benecke, Annick Lesne, and Marc-Thorsten Hütt *A network perspective on metabolic inconsistency*

Abstract

Integrating gene expression profiles with metabolic pathways under different experimental conditions is helpful for understanding how the conditions affect the coherence of these two layers of cellular organization. The GIMME algorithm proposed by Becker and Palsson [18] permits the incorporation of gene absence/presence patterns as experimentally determined constraints in constraint-based reconstruction analysis (COBRA). In this study we elaborate on the inconsistency measure resulting from GIMME, which quantifies the discrepancy of the provided data to the applied cellular objective. So far it has been mainly used for quality control assessment of the simulated flux distributions specific for a gene expression data set. In this study we compare the inconsistency to the metabolic coherence (MC), a measure which solely relies on the metabolic system's architecture, and determines the correspondence between gene expression data and the metabolic network. We find a surprisingly strong negative correlation between both measures indicating that the dynamical quantity obtained by GIMME identifies the same patterns as the purely topological quantity.

This observation in particular allows us to investigate the individual contributions of the inconsistency from a topological perspective. On this basis we are able to separate the specific contributions, i.e., those components of the inconsistency vector that bear valuable information about the dynamical system, from the unspecific contributions, which allow one to unravel gaps in the metabolic reconstruction.

Introduction

Genomic information allows compiling an inventory of an organism's enzymes and thus the subsequent reconstruction [40] and simulation of its metabolic system [66] using constraint-based modeling (CBM) techniques [95]. Compensating the lack of detailed information on the systems parameters, e.g., enzyme kinetics, gene regulation etc., CBM has proven to be a valuable tool for genome-scale system analysis. For example, flux balance analysis (FBA) [128] has been used to predict with high accuracy the lethality of gene deletions in unicellular organisms by taking only the metabolic system's stoichiometry, the assumption of optimal growth (implicit gene regulation), and a specified growth medium into account (see e.g. [36], for a study involving *Escherichia coli* or [34], for a study involving *Saccharomyces cerevisiae*).

Duarte et al. [35] published a genome-scale representation of human metabolism based on genomic, bibliographic, and biochemical information. In contrast to its predecessor models, being representations of unicellular organisms with specific media conditions, the following caveats play a role when it comes to the modeling of multicellular reconstructions in general and in particular for the human system: (i) it is difficult to define environmental conditions for a multicellular system, (ii) usually not much information is available about the cell-type specificity of human metabolic pathways, and (iii) cellular objectives, a prerequisite for flux balance analysis, are hard to define and validate.

The precision of CBM predictions increases with the availability and accuracy of the used constraints, as they help to narrow down the potential solution space to the biologically meaningful states. Thus, integrating experimental data can help overcome the previously mentioned limitations. For example, the integration of transcriptome data with metabolic reconstructions has shown to be useful for identifying subnetworks and regulatory interactions [32]. Nevertheless, the lack of significant correlations between gene expression and enzyme activities [30, 121], let alone RNA transcript levels [121], renders the incorporation of such data into CBM approaches difficult. This is especially true for higher multicellular eukaryotes, which are using a wide spectrum of post-translational regulatory mechanisms [30, 101].

Different approaches have been proposed for the incorporation of experimental data into CBM:

Akesson et al. [1] exploit the fact that under steady state conditions the absence of gene expression coincides with the corresponding protein unavailability, and thus inactivity of the corresponding reactions. Thus, the flux through an enzymatically catalyzed reaction is constrained to zero if the corresponding gene is not expressed in an experimental data set.

The GIMME (Gene Inactivity Moderated by Metabolism and Expression) algorithm proposed by Becker and Palsson [18] relaxes the rigid approach by Akesson et al. [1] by reinserting unexpressed reactions back into the system if a proposed cellular objective is not achieved. The sum over these reinserted fluxes is termed inconsistency (I) and is minimized during the GIMME optimization. The inconsistency I gives, on the one hand, an estimate of the quality of the computed flux distribution, and measures, on the other hand, the coherence of the objective and the experimental data. GIMME has been applied recently in an investigation of a joint model of human alveolar macrophage and *Mycobacterium falciparium* physiology [21].

Shlomi et al. [112] proposed an mixed-integer optimization problem formulation that allows the computation of tissue-specific, steady-state flux-distributions that match experimental data at hand as close as possible. In particular, it is not necessary to formulate a cellular objective function.

The E-flux method developed by Colijn et al. [28] uses experimental data directly as boundaries in the linear programming formulation. It has been successfully utilized for a comprehensive study of *Mycobacterium tuberculosis* mycolic acid metabolism predicting accurately the effects of a series of anti-TB drugs [28].

In this piece of research we will integrate human transcriptome data sets from healthy and tumorous adrenal gland tissues with the metabolic reconstruction Human Recon 1 [35]. We will use the GIMME algorithm [18] for this purpose because its inconsistency measure suits our approach of quantifying the discrepancy of the measured transcript levels to a given cellular objective, e.g., ATP or aldosterone production.

The optimization problem which is solved by GIMME can be formulated in the following way:

$$\begin{aligned} \text{Minimize} \quad & I = \sum_{j=1}^{n} p_j |v_j| \\ \text{subject to} \quad & \boldsymbol{S} \cdot \boldsymbol{v} = 0 \\ & \boldsymbol{v_{min}} < \boldsymbol{v} < \boldsymbol{v_{max}} \\ & v^{obj} \geq v_{max}^{obj} \cdot l. \end{aligned} \tag{5.1}$$

The inconsistency score I is, technically speaking, the sum of all fluxes going through unexpressed reactions weighted by the respective experimental data p_j. Furthermore, n is the number of reactions/fluxes v, $\boldsymbol{S}$ represents the stoichiometry of the system as a matrix, and v_{max}^{obj} is the maximal flux through the proposed objective reaction v^{obj} (v_{max}^{obj} is determined in a previous step by standard FBA without taking the

experimental data into account). The condition $v^{obj} \geq v^{obj}_{max} \cdot l$ forces the system to operate at or above some level l chosen from the interval $(0, 1]$. The norm $|v_j|$ can be omitted by using exclusively irreversible reactions. This can be achieved by replacing reversible reactions with pairs of irreversible reactions.

The weighting vector p_j is constructed in the following way:

$$p_j = \begin{cases} t - x_j & \text{if} \quad x_j < t \\ 0 & \text{if} \quad x_j \geq t \end{cases}, \tag{5.2}$$

where t is a threshold applied to the gene expression data x that classifies reactions as either expressed ($x_j \geq t$) or not expressed ($x_j < t$) using the gene expression data $\boldsymbol{x}$. Fluxes through expressed reactions are thus not minimized in eq. (5.1). Conversely, the usage (reinsertion) of fluxes through unexpressed reactions are weighted by the distance of the expression level from the threshold $t - x_j$ in eq. (5.1).

We will compare the inconsistency I to the metabolic coherence (MC) introduced in [117], which is a purely topological quantity that measures the fragmentation of effective networks. The coherence of metabolic network topology and gene expression patterns is quantified as follows (see also Methods): we map genes with $\boldsymbol{x} > t$ directly onto a metabolic gene network of human metabolism in order to extract effective subnetworks. Then we compute the ratio of connected nodes and overall nodes in the effective subnetwork. This ratio is then converted into a z-score, by using a random distribution of expression changes as a null model. This z-score is our *metabolic coherence* (MC), which measures the amount of network coherence between gene expression profiles and metabolic pathways.

Figure 5.1 shows a flow diagram that describes the structure and necessary steps of our comparative analysis. Based on this quantitative comparison we will attempt a topological characterization of the individual contributions to I and show that valuable information can be extracted from them.

Results

Inconsistency uncovers two types of metabolic behavior

Figure 5.2 shows the distribution of inconsistency values for the control and adenoma transcript profiles. Maximal aldosterone production was used as the cellular objective function v^{obj} and a minimal medium composition containing glucose and glycerol, as well as a collection of amino acids and fatty acids, was implemented using the appropriate boundaries on the their respective exchange reactions (for further details see Methods). The histogram in Figure 5.2 uncovers a bimodal distribution of adenoma inconsistency values: it consists of a group of adenomas exhibiting low-

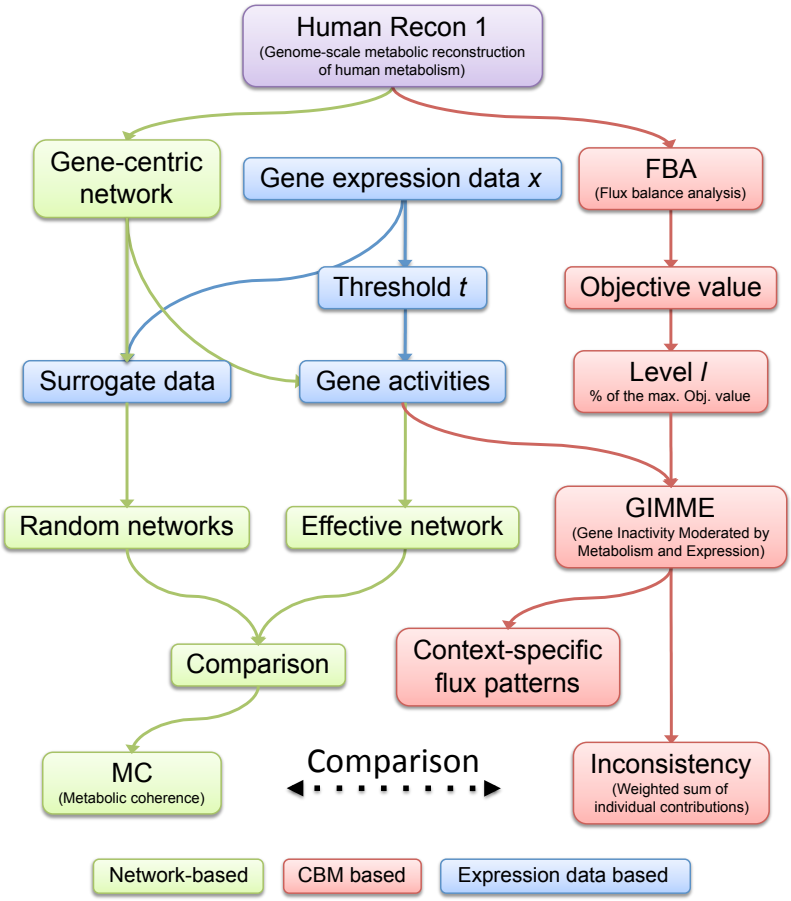


Figure 5.1: A schematic figure explaining methodological approach behind our comparison of metabolic coherence (MC) and inconsistency I.

er(higher) inconsistencies than the average $\langle I \rangle$ of the control group. We term them high/low inconsistency group respectively (HIG and LIG).

Comparison of metabolic coherence and inconsistency

Figure 5.3a shows a scatterplot of the metabolic coherence and inconsistency values for all 69 expression profiles (58 adenomas + 11 controls), using the reference medium and the threshold $t = 2.0$. It is obvious that both measures are strongly anti-correlated (Pearson's product correlation coefficient $r = -0.64$; Spearman's rank correlation coefficient $\rho = -0.67$)

Figure 5.3b shows the dependency of the correlation on the threshold t that is applied to the data to distinguish between expressed and not expressed metabolic genes. The strongest negative correlation appears for parameter ranges that fit our

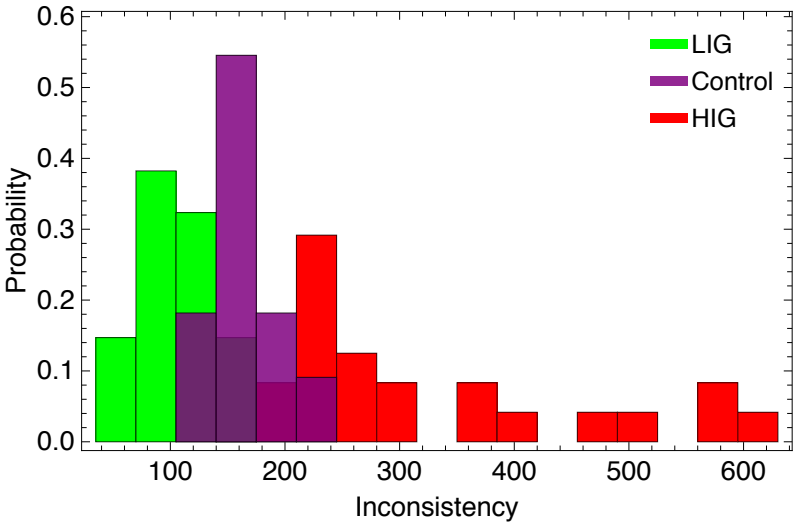


Figure 5.2: Distributions of inconsistency scores for the adenoma and control data with aldosterone production as objective function.

statistical understanding of the raw signal distribution (see Figure C.1). Figure 5.3c shows that there is no clear dependence on the level parameter l that is used to enforce a certain flux through the stated objective v_{obj}.

In Figure 5.3d the dependency of the correlation between inconsistency and the MC on the chosen growth medium is shown. The distribution of correlation coefficients is narrow (between -0.75 to -0.4 for r and -0.73 to -0.35 for ρ), regardless of which correlation measure is considered. This indicates that both inconsistency and MC and their correlation seem not be strongly dependent on the environmental conditions provided. Furthermore, the correlation values obtained for the reference medium (i.e. $r = -0.64$ and $\rho = -0.68$, see also above) seem to be originated on the left tail of the distributions, suggesting a rather high correspondence with the *in vivo* situation.

Individual contributions to the inconsistency

The correlation between metabolic coherence and inconsistency suggests a connection between both measures, and thus the possibility of interpreting the inconsistency values from the perspective of network topology. In order to investigate this point, we will decompose the inconsistency value into a vector of individual contributions, i.e., reactions that have been reinserted during the optimization procedure in order to achieve the targeted flux-level of the objective function. We further define the contribution strength of a reaction as the number of contributions it makes to the inconsistencies of a data set divided by the size of the respective data set.

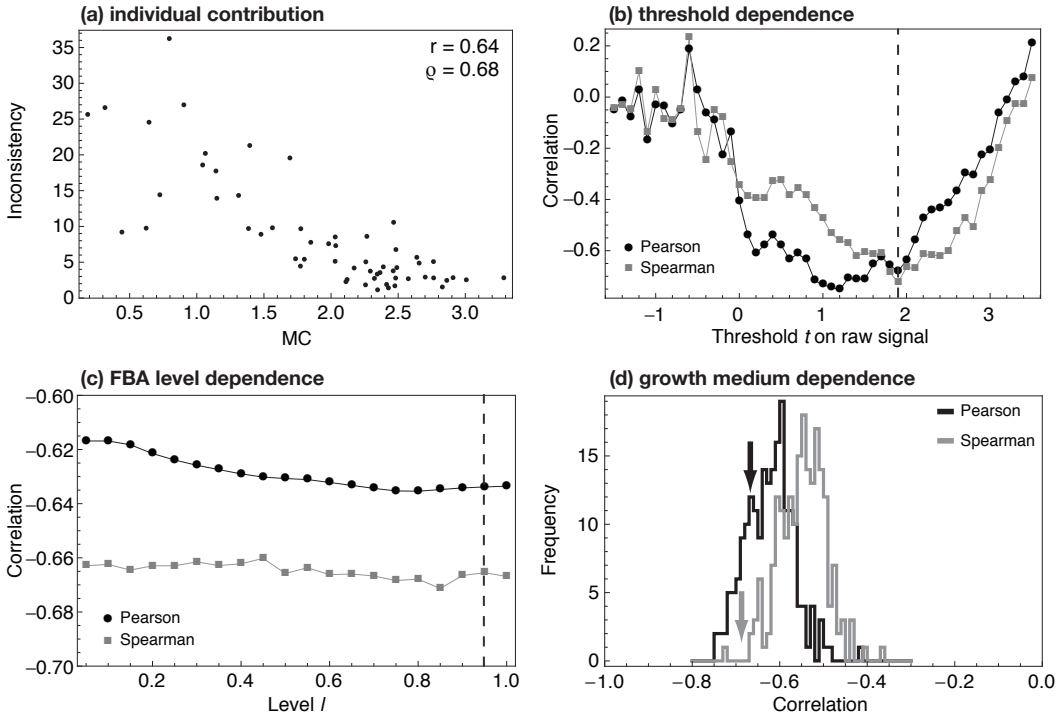


Figure 5.3: (a) The ATP-production inconsistency values are plotted against the MC of 69 tumor and control data sets. A clear negative correlation is visible (Pearson's product-moment correlation coefficient $r = -0.64$, and Spearman's rank correlation coefficient $\rho = -0.68$ ($t = 1.9$; $l = 0.95$). (b) Dependency of the correlation on the threshold parameter ($l = 0.95$). (c) Dependency of the correlation on the level parameter ($t = 1.9$). (d) Medium dependency of negative correlation strength. Both Spearman's rank correlation coefficient as well as Pearson's correlation where computed for the MC and the inconsistency for 100 random growth media ($t = 2.$; $l = 0.95$). Dashed lines in (b) and (c) indicate the parameters that have been used in (a). Arrows in (d) indicate the correlation values found in (a).

Figure 5.4 displays inconsistency contributions and flux patterns for control, HIG and LIG on the carbohydrate metabolism pathways of *E. coli*. Striking differences between the control and the adenoma group become visible in this overview, e.g., the pentose-phosphate pathway seems to be activated only in the HIG and LIG. Furthermore, the control pathway map lays out a rather consistent pattern of flux activities. Though, a few exceptions arise: pyruvate dehydrogenase (PDHm), the major entry point to the TCA cycle, and to some lesser extent hexokinase (HEX1) and the pyruvate transport from cytosol to mitochondrium (PYRt2m), exhibit high contribution strengths. It is intriguing that the contribution strengths for these particular reactions is diminishing small in the LIG case, which indicates an elevated energy metabolism for those adenomas. The HIG, on the other hand, shows a large number of reactions with elevated contributions strengths homogeneously distributed over the whole map, which indicates a significantly reduced energy metabolism.

How do the contribution strengths vary between control, LIG and HIG on a systems-wide level? Figure 5.5 shows the contributions for a subset of all contributing reactions (see Figure C.6 for the complete set of contributions). The contributing reactions have been sorted according to their contribution strengths in the control group and at equal strength by the overall contributions. Having already seen a few examples for differentially contributing reactions in the carbohydrate pathways, it becomes ever more clear that there are many more to discover.

On the other hand, a group of reactions with very high contribution strengths seems to contribute non-specifically and independently from the gene expression data. In the following, we want to elaborate on this set of reactions, and will use certain categories and topological markers to characterize them.

The following circumstances can lead to non-specific contributions to the inconsistency vector:

1. A reaction is *expressed in vivo* but the measured gene expression intensity falls below the threshold t under most or all experimental conditions (*just below threshold*). This is a consequence of the rigid application of a universal threshold. Topologically, these contributions often disrupt a chain of otherwise expressed reactions (*chain disruptor*).

2. A reaction is *expressed in vivo* but, e.g., wrong GPR associations, missing isozymes, wrong gene annotations, erroneous data etc., make it invisible for the analysis. Again, these artifacts are often characterized by an interrupted chain of expressed reactions (*chain disruptor*).

3. The reaction is *not expressed* but it has to be utilized by GIMME due to the following reasons:

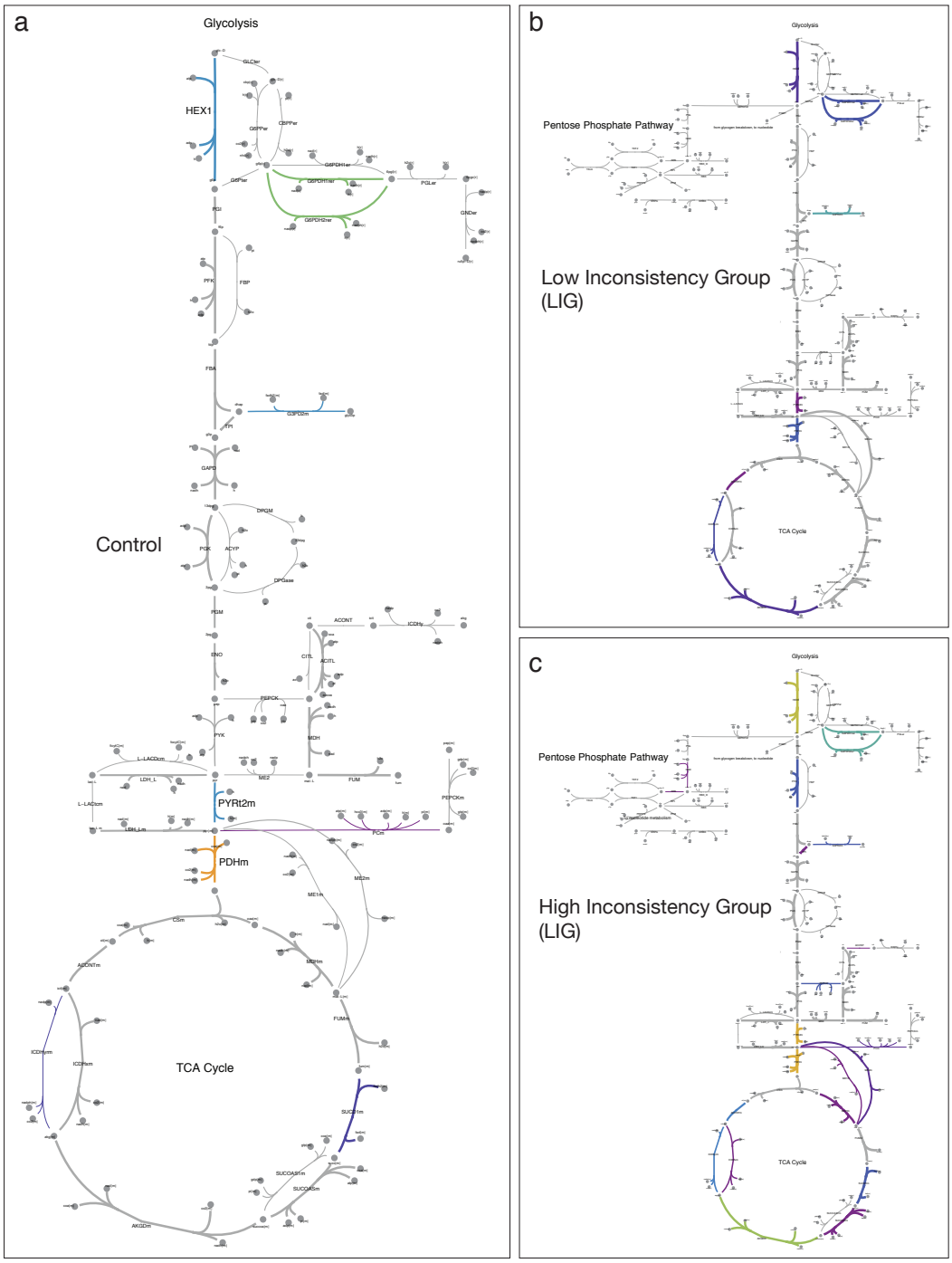


Figure 5.4: Inconsistency contributions to carbohydrate metabolism. The maps depict the usage patterns and inconsistency contributions for the control and the high and low inconsistency groups (HIG and LIG). The thickness and color of a reaction edge correspond to the usage frequency and the contribution strength, respectively. The pathway maps have been obtained from the BIGG database [105]. Please refer to the electronic version of this manuscript for a high resolution depiction of the effective pathways.

Figure 5.5: Inconsistency contributions. (a) Overall contributions to the inconsistency for all adenoma and control samples, (b) the control group, and the adenoma tumor samples showing (c) lower (LIG) and (d) higher (HIG) inconsistencies in comparison to the control group. Black arrows indicate differentially contributing reactions and the gray box indicate the group of unspecific reaction contributions. Both of them are covered in Table 5.1. Only a subset of all contributing reactions is shown due to space limitations. The complete diagram is shown in Figure C.6. Please refer to the electronic version of this manuscript for a high resolution depiction of this figure.

a) The stated objective function does not reflect the situation present in the cell. Defining the objective functions as the output of the system, these reactions contributions should often lie close to it (*close to output layer*).

b) The chosen media composition does not reflect the *in vivo* environment in which the experimental data has been obtained. The preliminary FBA step in GIMME is naive about the *in vivo* medium composition and uses everything provided and suitable for the maximization of the objective function. As GIMME enforces a certain achievement of the objective flux predicted by FBA, many of the transport reactions used by FBA will also be used by GIMME. Topologically, these reaction contributions are characterized by lying close to the provided medium components (*close to input layer*).

c) Too many missing gene-protein-reaction associations (GPR), either due to non-enzymatic reaction steps or knowledge gaps, before and after the contributing reaction can lead to wrongly activated paths, as missing GPR information is not punished by GIMME (*invisible path*).

d) Alternative *expressed* routes to the objective function are available *in vivo* but are not covered by the metabolic reconstruction. This leads to reaction contributions that are characterized by producing essential precursors for the objective function, and thus constitute bottlenecks in the system (*bottleneck*).

Table 5.1 lists topological and biological classifications for the 11 unspecific contributions as well as 8 selected differentially contributing reactions (see Figure 5.5). The topological characterization from the enumeration above have also been applied to the specific contributions.

Conclusions

Using GIMME [18] we provided context for a series of gene expression profiles obtained from adrenal gland adenomas and healthy tissues. Employing the inconsistency measure as tool for measuring the discrepancy of the provided data and aldosterone production, a suitable objective for these tissues, we were able to detect two groups of adenomas, characterized by distinct physiological behaviors, i.e., LIG and HIG.

Comparing the inconsistencies with the metabolic coherence exhibited a connection of both measures, leading us to investigate the characteristics of the individual contributions on the reaction level. Comparing the contribution strengths of individual reactions among the different sample categories (control, LIG, HIG) revealed

Table 5.1: Classification of contributions to the inconsistency vector. Reactions are sorted according to their contribution strength and follow the same order as depicted in Figure 5.5. All unspecific and a few selected specific contributions are provided. Topological characterizations as well as biological interpretations are given. References point to positions on the electronically provided pathway maps (see Appendix C).

Contributor	Category	Topological class	Biological interpretation	Reference
AATAi†	unspecific	bottleneck; invisible path	*2-Aminoadipate transaminase*; one out of two 2-oxoadipate producing reactions; missing GPR assoc. in all precursors.	SIaa, B1
PROD2	unspecific	chain disruptor; close to input layer	*Proline dehydrogenase*; participates in a cycle that converts nadh to fadh2 (Figure C.4 and Appendix C); not expressed (Figure C.7b).	SIaa, D3
DPMVDx	unspecific	chain disruptor	*Diphosphomevalonate decarboxylase*; essential step in the cholesterol biosynthesis pathway; not expressed; wrong or missing GPR assoc.? (Figure C.7c).	SIlip, B5
OIVD(1,2,3)m	unspecific	chain disruptor; close to input layer	*2-Oxoisovalerate dehydrogenase*; necessary for leucine(valine, isoleucine) processing; expression just below threshold (Figures C.7d, e and g).	SIaa, D2–E2
2OXOADPTm†	unspecific	bottleneck; invisible path	*2-Oxoadipate shuttle (cytosol/mitochondria)*; involved in the 2-oxoadipate producing pathway; missing GPR assoc. in precursor reactions.	SIaa, A2
GLYK	unspecific	close to input layer	*Glycerol kinase*; not expressed in control and LIG, indicating that glycerol (provided in the *in silico* medium) might not be available as a *in vivo* medium component; slightly elevated expression levels in LIG (see Figure C.7i).	SIlip, E5
MACACI‡	unspecific	chain disruptor; close to input layer	*Maleylacetoacetate isomerase*; expression levels just below threshold (see Figure C.7j).	SIaa, B5
PHETHPTOX2	unspecific	close to input layer	*Phenylalanine 4-monooxygenase*; converts phenylalanine (provided in the *in silico* medium) into tyrosine (not provided in the *in silico* medium); the high unspecific contribution strength indicates that tyrosine might be available as an *in vivo* medium component.	SIaa, A5
COAtm	unspecific	—	*CoA transporter (cytosol/mitochondria)*; expression just below threshold (see Figure C.7i).	no map
34HPPOR‡	unspecific	chain disruptor; close to input layer	*4-Hydroxyphenylpyruvate dioxygenase*; involed in tryosine to fumarate and acetoacetate conversion; not expressed (see Figure C.7m).	SIaa, B5
HGNTOR‡	unspecific	chain disruptor; close to input layer	*Homogentisate 1,2-dioxygenase*; involed in tryosine to fumarate and acetoacetate conversion; not expressed (see Figure C.7n).	SIaa, B5
FUMtm	unspecific	—	*Fumarate transport (cytosol/mitochondria)*; expression just below threshold (see Figure C.7o).	no map
GLUTCOADHm†	specific	chain disruptor	*Glutaryl-CoA dehydrogenase*; involved in the 2-oxoadipate pathway (see Figure C.3); elevated expression levels in LIG (see Figure C.8a).	SIaa, A2
PGCD	specific	bottleneck	*Phosphoglycerate dehydrogenase*; expression is just below the threshold in the control, decreased expression levels in LIG an HIG (see Figure C.8b).	SIaa, E5
PDHm	specific	bottleneck	*Pyruvate dehydrogenase*; entry point to the TCA cycle; elevated expression levels in LIG (see Figure C.8c).	SIcarb, C3
MMEm	specific	chain disruptor	*Methylmalonyl-CoA epimerase*; involved in isoleucine degradation; slightly elevated expression levels in LIG (see Figure C.8d).	SIaa, D1–E1
MEVK1x	specific	chain disruptor	*Mevalonate kinase*; an essential step in cholesterol biosynthesis; decreased expression levels in LIG and HIG (see Figure C.8e).	SIlip, B5
G6PDH(1,2)rer	specific	topo	*Glucose-6-phosphate dehydrogenase*; slightly elevated expression levels in LIG (see Figure C.8f).	SIcarb, C4–D4
DHCR71r	specific	chain disruptor; close to output	*7-Dehydrocholesterol reductase*; involved in cholesterol biosynthesis; slightly elevated expression levels in LIG (see Figure C.8g).	SIlip, A4
HEX1	specific	chain disruptor; close to input layer	*Hexokinase*; first step in glycolysis; slightly elevated expression levels in LIG (see Figure C.8h).	SIcarb, C4–C5
SQLEr	specific	chain disruptor	*Squalene epoxidase*; decreased expression levels in LIG and HIG (see Figure C.8i).	SIlip, A5

† 2-Oxoadipate issue; see Figure C.2, Figure C.3 and Appendix C ‡ Tyrosine path to fumarate and acetoacetate issue; see Figure C.5 and Appendix C

a group of unspecific contributors to the inconsistency, as well as a group of reactions differentially contributing in a specific fashion.

Topological markers developed for the characterization of both, specific and unspecific contributions, have proven be valuable for a thorough understanding of the context-specific flux-activity results.

Material and Methods

Model of human metabolism

All flux balance simulations were conducted using human Recon 1, a genome-scale compartmentalized representation of human metabolism [35], which is available in SBML [58] format via the BIGG database [105].

Gene expression data

Logarithmized transcript levels from 58 adenomas and 11 control tissue samples were mapped onto the GPR (gene-protein-reaction) associations included in the Human Recon 1 model. Therefor, it was necessary to replace logical AND and OR by min and max functions, respectively, following the protocol described in [18].

Context-specific flux balance analysis

Context-specific flux balance analysis of human expression data was conducted using the GIMME algorithm as described in [18] and in the introduction to this work. ATP-production was implemented as a cellular objective by introducing an artificial reaction that consumes cytosolic ATP. The aldosterone objective was implemented as the maximization of flux through aldosterone synthase (P45011B21m). The pathway to aldosterone was initially blocked in the metabolic reconstruction. Further analysis revealed 4-Methylpentanal as a dead-end metabolite inhibiting steady-state flux to the aldosterone synthase reaction. The introduction of an artificial drain for 4-Methylpentanal restored the functionality of the whole pathway. Furthermore, the same conservative approach was chosen regarding missing GPR: reactions without GPR associations were assumed to be expressed, i.e., having expression values above t. The aldosterone objective and the parameters $t = 2$ and $l = 0.8$ were used throughout the study, if not stated otherwise.

Growth media

The growth medium was defined as in [113] (see Table C.1). It contains both glucose and glycerol as carbon sources, the amino acids L-arginine, L-histidine, L-isoleucine,

L-leucine, L-lysine, L-Methionine, L-phenylalanine, L-threonine and L-tryptophane, as well as the fatty acids palmitic and linoleic acid. Aerobic conditions were assumed by leaving oxygen consumption unconstrained. Random media conditions were constructed by picking approximately the same number of exchange reactions as in the reference medium randomly and assigning random upper and lower boundaries in the intervals $[-20, 0]$ and $[0, 20]$ to them. Oxygen, protons, sulfate, phosphate, water were assumed to be always available. In case of the random media sampling, inconsistency values I have been normalized by the objective function's flux in order to make them comparable.

Metabolic coherence

The metabolic coherence (MC) was computed as described in Sonnenschein et al. [117]. We mapped the set of transcribed genes (genes with expression values above a certain threshold are considered as expressed) directly onto a metabolic gene network representation of human metabolism in order to extract effective networks (i.e., the network spanned by the significant expression changes). Before constructing the gene network out of the bipartite representation, overly abundant currency metabolites, e.g., ATP, H_2O, NADH etc., have been excluded by removal of 4% of the highest connected compounds in the network [73]. Then we computed the ratio of connected nodes to overall nodes in the effective network. Using a random distribution of expression changes as a null model, we converted this ratio into a z-score, the *metabolic coherence* (MC).

Chapter 6

Metabolic variability in adrenal gland tumors

This chapter presents contributions to the following work in progress [22]:

> Sheerazed Boulkroun, José-Felipe Golib Dzib, Nikolaus Sonnenschein, Hélène Leger, Benoit Samson-Couterie, Arndt Benecke, Marc-Thorsten Hütt, and Maria-Christina Zennaro *Using flux balance analysis to link gene expression with pathological profiles of a large cohort of human primary aldosteronism patients.*

Abstract

As provided by our collaborators:
We present a transcriptome analysis coupled with methods from constraint-based modeling of human metabolism on a cohort of 134 human Primary Aldosteronism (PAL) patients. Very little is known about the primary causes that determine PAL, consisting in an autonomous overproduction of the mineralocorticoid hormone aldosterone from the adrenal gland. This is the most frequent cause of secondary hypertension (high blood pressure) in humans [23]. The aim of our work is to identify the activation of particular pathways or transcriptional cascades, via qualitative (mutations) or quantitative (expression) molecular changes that may be responsible for the development of PAL. We obtained 183 transcriptome profiles using AB1700 microarray technology, which has shown to provide highly sensitive gene expression

measurements. These profiles were correlated with clinical annotations to identify biologically relevant molecular traits that may explain the differences among the patients' pathological profiles.

Furthermore, this data was integrated with a genome-scale reconstruction of human metabolism [35] in order to predict tumor-specific metabolic pathways using the GIMME method [18]. GIMME (Gene Inactivity Moderated by Metabolism and Expression) exploits the gene expression data as a proxy for enzyme availability and computes flux activities that maximize a stated metabolic objective, e.g., ATP or aldosterone production. Our results herald a close relationship between the inferred metabolic activity and the molecular morphopathology of PAL and furthermore reveal an unexpected diversity of metabolic pathway usage.

Energy and steroid metabolism are strongly coupled

Plotting the inconsistencies obtained through the aldosterone and ATP objective in a scatter plot reveals a strong positive correlation for all samples (see Figure 6.1). Does

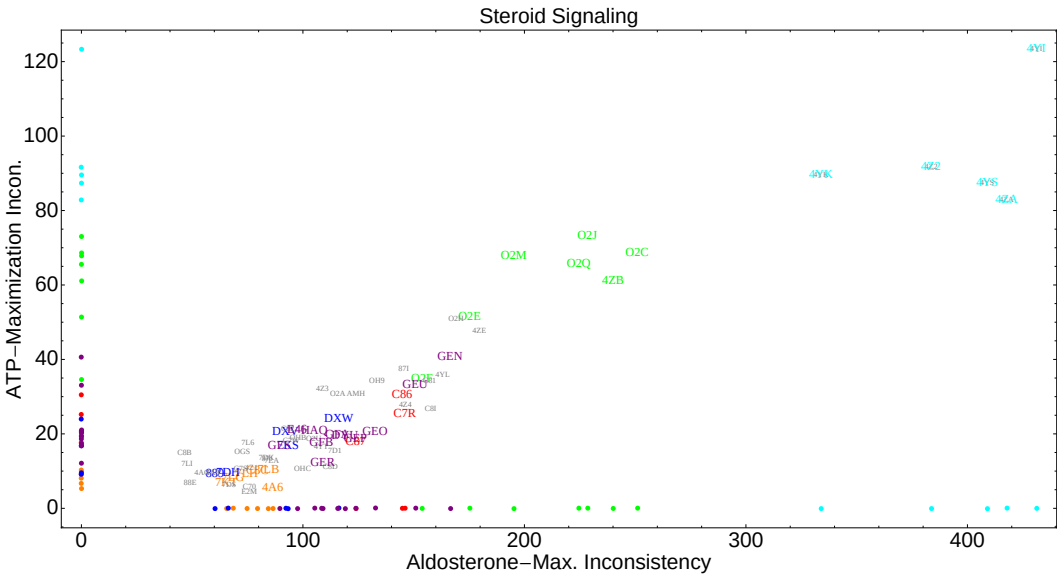


Figure 6.1: The inconsistency values obtained through the ATP and aldosterone objectives show a strong correlation. Purple labels indicate control samples and red, green, blue, orange labels represent membership in clusters, which have been obtained by traditional cluster analysis on the transcript profiles of genes involved in steroid biosynthesis.

this correlation allow for some biological interpretation? Is it indicating that tumors with reduced aldosterone production capabilities simultaneously show a reduced energy metabolism? Or is it simply a consequence of the simulation procedure?

In order to answer this question we investigate the relationship of ATP and aldosterone inconsistency in a different set of expression data. We chose the whole genome transcriptome survey provided by Dezso et al. [33] as a suitable reference data set for this purpose. It covers 32 human tissue types with three replicates per tissue summing up to a total of 96 microarray experiments (GEO accession GSE7905).

Figure 6.2 shows the inconsistency profiles of the 96 tissue samples together with the previously shown inconsistencies of the adenoma transcript profiles (see Figure 6.1). The analysis of the 96 human tissue samples reveals a general connection between the two objectives. Nevertheless, the strongly diverging slopes of the two inconsistency profiles reveal for the adrenal gland samples a significantly stronger agreement with the aldosterone objective.

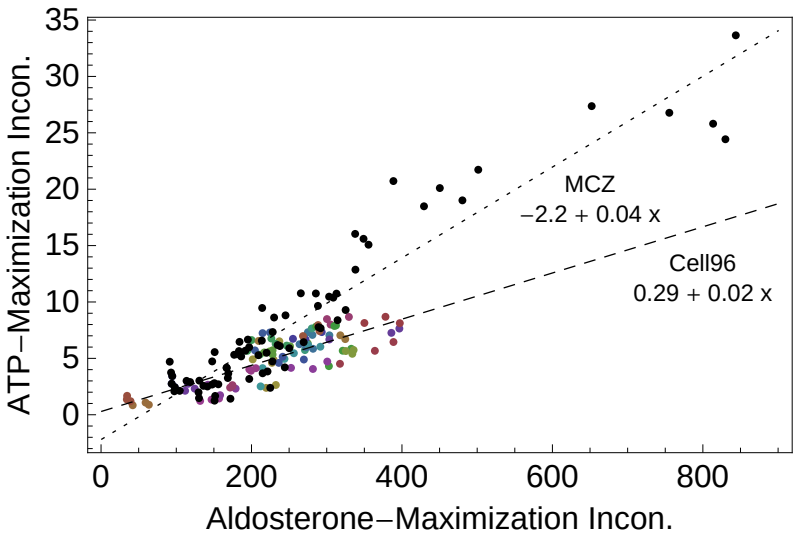


Figure 6.2: The strong correlation between the ATP and aldosterone objective is also found for the Cell96 data set but the slopes of a linear least-squares fit are different.

Context-specific flux profiles reveals a strong variability between the adenoma samples

Context-specific analysis not only provides the inconsistency measure as an indicator for the disagreement of the transcript data with the stated objective, but furthermore produces flux-activities that are in closest agreement with the provided data [18]. Thus, instead of comparing the samples using their respective scalar inconsistency values, these flux profiles allow us to increase the resolution of our comparisons even further.

Figure 6.3 shows the result of a cluster analysis of a distance matrix that represents all pairwise comparisons of context-specific tumor and control flux-activities. The distance between two flux-activity profiles was defined as:

$$D = \frac{|f_1 \cap f_2|}{min(|f_1|, |f_2|)}, \tag{6.1}$$

where f_1 and f_2 are the compared flux-activities. The distance D ranges in the interval $0 \leq D \leq 1$, where $D = 1$ if the smaller flux set is a complete subset of the other set and $D = 0$ if $|f_1| \cap |f_2| = \varnothing$.

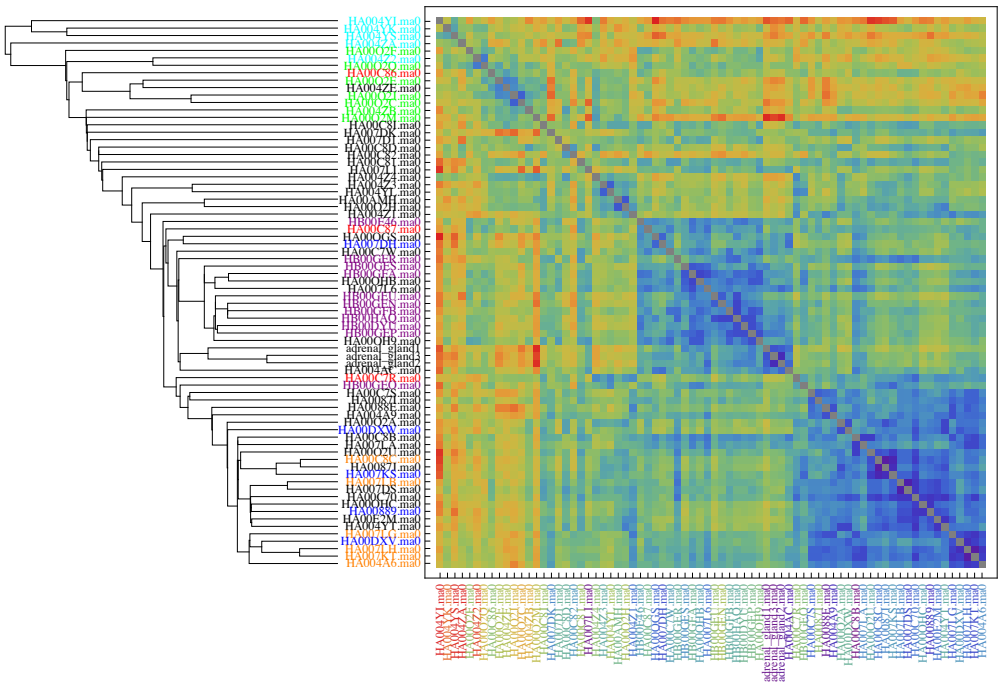


Figure 6.3: Clusters of context-specific flux distributions obtained by GIMME using the aldosterone objective. The rows and columns of the distance matrix have been sorted according to the dendrogram obtained by the cluster analysis (distance metric defined in eq. 6.1; hierarchical clusters computed by *Ward's* minimum variance method). Purple dendrogram leaf labels indicate the control group and red, green, blue, orange labels represent membership in clusters, which have been obtained by traditional cluster analysis using the transcript profiles of genes that are involved in the steroid biosynthesis pathway. The color of the sample labels on the horizontal axis of the grid plot code for the inconsistencies, where red indicates a high and purple a low inconsistency, respectively.

The dendrogram as well as the colored and sorted distance matrix in Figure 6.3 show a series of interesting features:

(i) The control group (purple) clusters strongly together, (ii) the three adrenal gland samples from Dezso et al. [33] behave quite similar to the control group, (iii) a group of adenomas, showing consistently lower inconsistencies in comparison to the control group, cluster strongly together and are quite dissimilar to the control (lower right of the colored grid), (iv) a group of adenomas, showing consistently higher inconsistencies in comparison to the control, do not cluster together, and (v) clusters which have been obtained by traditional cluster analysis using the transcript profiles of genes that are involved in the steroid biosynthesis pathway, colored in orange, blue, red, green and cyan, respectively, seem to cluster mostly together.

Features (i) and (ii) basically tell us that there exists some kind of normal mode of operation, which is quite distinct to (iii), a group of tumor cell that seem to have a higher capability of producing aldosterone. Furthermore, the lack of a clear pattern in (iv) seems to be an effect of the increasing inconsistency: transcript samples, which are associated with sufficiently high inconsistencies, or in other words, to many individual contributions, are not constraining the system in a systematic fashion, leading to arbitrary patterns of flux-activities. The samples in (iv) definitely do not match the aldosterone production objective leaving us with the question which objective they might match. The few exceptions to (v) could be interesting, as they provide evidence for a differential behavior on a systems-wide level, e.g., the red cluster presents a group of three samples that behave quite similar on the pathway-level of steroid biosynthesis but are very disjunct on the overall metabolic level.

Methods

See Methods in Chapter 5.

Chapter 7

Conclusions and outlook

Understanding dynamic processes on networks, particularly in an evolutionary context, has over the last few years become an important endeavor in the field of network and systems biology. Particularly gene regulatory networks and metabolic networks have received a large amount of attention over the last years. Currently, systems biology descriptions focus mostly on these subsystems separately, although in cellular function they are heavily interlinked. In this work we have analyzed both systems separately, as well as in conjunction.

We have shown in Chapter 2 that two categories of genes—with ("regulatory") and without regulatory information available ("isolated")—unmix on the chromosome of *Escherichia coli*. In the light of the findings by Marr et al. [81], who established the concepts of *digital*[1] and *analog*[2] control of transcriptional regulation, we argue that the second category of "isolated" genes is regulated by the latter. In order to support this claim we furthermore investigated the associations of the two categories of genes with highly expressed and transcriptionally silenced domains in the chromosome, being characterized by a high protein abundance [130]. We used these protein occupancy domains as proxies for the *analog* type of control, and indeed found in comparison to the "regulatory" genes a stronger association of the "isolated" genes with these domains. The success of the point-process statistical methods [62], which, to our knowledge, have been applied for the first time to the spatial organization of genes, suggests an extension to other categories of genes and chromosomes of other organisms.

[1] *Digital* in the sense of the almost Boolean mode of activatory and inhibitory actions performed by transcription factors.

[2] *Analog* in the sense of the rather continues control of chromosomal architecture.

In Chapter 3 we extended our studies on control types in a data-driven fashion. Using a network model of *E. coli* metabolism, we classified gene expression changes that have been obtained under variation of chromosomal supercoiling, in wild type *E. coli* as wells as in a series of mutants lacking abundant structural proteins, by quantifying the coherence of the data to the network topology. We found a strong coherence in the wild type patterns, indicating that the gene expression changes induced by chromosomal rearrangements are indeed meaningful in a metabolic context. Furthermore, the significant reduction of coherence observed for the mutants showed that the deleted structural proteins are responsible for translating the chromosomal topology into meaningful gene expression patterns. Furthermore, we were able to exclude the possibility that the transcriptional regulatory network (TRN) acts as the mediator for the coherence found in the wild type case by evaluating the consistency of the signs of the expression changes with the logic of the TRN. These results provide strong evidence that the expression changes induced by *analog* control are indeed a major organizational factor of bacterial physiology.

On the experimental side, with next-generation sequencing data becoming affordable, it would be nice to reproduce and expand the analysis of supercoiling induced expression changes. Furthermore, by using more quantitative data it would be possible to apply constraint-based modeling techniques (see also Chapter 5 and 6) in order to quantify the reduced coherence in more detail and also compare it our previous results. This would help in shifting the analysis from a global statistical treatment to a more fine grained perspective, and thus to potential new biological insights.

In Chapter 4 we elaborated on topological markers of reaction essentiality, in particular, medium dependent and independent essentiality as well as inessentiality of reactions. Our method of using multiple topological labels for investigating the relation between network properties and system properties and in this way identifying sub-categories emanating from different topological features may be helpful in a broad range of contexts in systems biology. In the case of metabolism, we observe that some combinations of reaction categories show a substantially higher accuracy in predicting essentiality, suggesting different types of essential reactions.

In Chapter 6 and 5 we again integrated microarray expression data with a metabolic network, this time based on a genome-scale reconstruction of human metabolism [35]. RNA transcript levels of healthy and cancerous adrenal glands were either classified using the *metabolic coherence* developed in Chapter 3, representing a purely topological measure, or with a constraint-based modeling approach [18], quantifying the inconsistency of the data and the cellular objective of aldosterone production.

Furthermore, the strong connection found in the comparison of the two approaches allowed us to interpret the metabolic inconsistency score obtained with GIMME from a network perspective. This inconsistency score basically constitutes the sum over all fluxes going through unexpressed reactions. We considered it as a vector of individual contributions, which can be classified according to their distribution in

the metabolic network. In particular, we observed an unspecific set of reactions contributing due to the limitations of the method (gaps in the metabolic reconstruction, unsuitable environmental conditions etc.) and a specific set of informative contributions. It turned out, that on the one hand, the specific contributions cast light on an unforeseen diversity of alterations in the physiology of adrenal gland adenomas and, on the other hand, the unspecific contributions could provide good entry points for the iterative refinement of the metabolic reconstruction.

Appendix A

Supplementary Information for "Analog regulation of metabolic demand"

Metabolic network representations

The strength of graph theory is that it can represent a complex system in a unified formal language of nodes and links. Examples of graph-theoretical analyses of metabolic systems are [47] and [64]. Essentially, the pattern of zero and non-zero entries of the stoichiometric matrix defines a graph representation of the metabolic system. The level of information conveyed by looking at metabolism from a graph-theoretical perspective is still subject to constant scrutiny [83].

A suitable approach for analyzing the correspondence between expression changes and metabolism and thus quantifying metabolic coherence is the application of the tools developed for the effective TRNs [81] to a gene-centric representation of metabolism, where the nodes are metabolic genes and a link is drawn between two genes, if the associated reactions share a common metabolite [73]. This representation is the gene-centric variant of one of the standard projections of a bi-partite graph representation of metabolism, where both, metabolites and reactions serve as node sets (see, e.g., 3, for a discussion of these and other representations). Additional forms of representation include metabolite-, reaction-, enzyme- or gene-centric views. The metabolite-centric view represents the interconversion possibilities of the different substrates, whereas other views concentrate more on the processes (reactions), pro-

tein (enzymes) and genomic (genes) levels respectively. We chose the gene-centric view for our analysis as it allowed us a direct comparison of expression patterns with metabolic pathways.

Reaction to gene mappings

Imagine the following scenario: a reaction is catalyzed by a single enzyme (no isozymes involved), encoded by a unique gene (no enzyme complexes involved). With this simple scheme in mind one might conclude that reaction-, enzyme- and gene-centric representations are redundant. However, most of the time reactions and their associated enzymes and genes are not interchangeable. Figure A.2 visualizes the amount of multiplicities between the reaction- and gene-level. The columns of the colored grid represent the absolute number of genes per reaction pair. The rows represent the number of unique genes. So the number of consecutive reaction pairs sharing a single gene can be found in column 2, row 1, as both reactions are associated with a single gene and it happens that it is the same for both. In contrast, the previously described simple scheme can be found in column 2, row 2. In order to assess the impact of these ambiguities on our results we constructed other graphs in addition to our gene-networks by applying the following rules: (1) the removal of reaction pairs lying beside the diagonal of the grid (see Figure A.2) excludes situations where a single or multiple genes are involved with both reactions; (2) taking into account only reaction pairs fulfilling the condition of the second column and row, thus effectively excluding enzyme complexes and the situations described in (1).

Currency metabolites

Another problem emerges through highly connected compounds, which have been termed current or currency metabolites in the past [59, 77]. They have caused reports of questionable average path lengths [10, 64] as they represent unrealistic shortcuts obscuring the essential pathway structures that have been assembled by biochemists over the last century.

For example Figure A.3 demonstrates the huge impact of *i*AF1260 metabolites on our MC results by showing the z-score pattern for the untreated *i*AF1260 network. All scores are basically below or a little above 1 and such no significant coherence could be measured.

However, the KEGG and EcoCyc data sets provide currency metabolite free representations through their human readable pathway maps (in contrast to their complete reaction databases). For *i*AF1260 [39] we employed on the one hand a threshold heuristic to remove a certain percentage (i.e., 4 % for the results shown in the main

article) of the most highly connected metabolites as described in [73], and on the other hand a manual curation of the network where currency metabolites were removed on a reaction to reaction basis, i.e., the approach described in Ma and Zeng [77]. Figure 3.3C (in the main text) shows the MC result for the manually curated network for comparison with the untreated one (see Figure A.3). Figure A.4 shows the dependency of the MC on the percentage threshold.

Constraint-based modeling

For a metabolic system consisting of N reactions and M compounds the linear programming (LP) formulation of FBA can generally be stated as follows:

$$\begin{array}{lll} \text{Maximize} & Z = \mathbf{c} \cdot \mathbf{v} & \\ \text{subject to} & \sum_{j=1}^{N} S_{ij} v_j, & i = 1, \ldots, M \\ & v_j^{min} < v_j < v_j^{max}, & j = 1, \ldots, N \\ & v_j^{(m,s)} < v_j^{(t)} < v_j^{(m,u)}, & j = 1, \ldots, N^{(t)}, \end{array} \tag{A.1}$$

where v is a vector of reaction fluxes constrained by the stated boundary conditions, S is a matrix storing the stoichiometric information of the system (i.e. S_{ij} is the stoichiometric coefficient of metabolite i in reaction j) and Z denotes the objective to be maximized represented by a linear combination of fluxes v_j and objective coefficients c_j. Here, $v^{(t)}$ denotes a transport reaction, i.e. a reaction either secreting metabolites from the system or taking them up. The quantity $N^{(t)}$ is the number of transport reactions. As an approximation to a rich medium condition we allowed for every available transport reaction $v^{(t)}$ unlimited secretion $v^{(m,s)} = -\infty$ and $v^{(m,u)} = 20$ [in units of $mmol/g \cdot dw \cdot h$] as an arbitrary upper bound to influx. With the exception of $v^{(t)}$, all reversible reactions were treated as two distinct irreversible reactions. Maximization of biomass production [39] and simultaneous minimization of all other fluxes was used as Z in order to avoid accumulation of flux in cycles. As all constraints are linear and the solution space is convex, a global maximum can always be found using linear programming (assuming the problem is well defined and not unbounded), though multiple global optima cannot be excluded [95].

Digital control and TRN consistency

The *E. coli* transcriptional regulatory network was obtained from RegulonDB [41] (version 6.4). Only links between transcription factors (regulators) and metabolic genes (as found in the EcoCyc network) were considered for the digital control and TRN consistency analysis. Digital control was measured in form of the *digital CTC*,

as described in [81], with the exception that the ratio of connected nodes to overall nodes was used instead of the ratio of connected to isolated nodes. Methodologically, this method is similar to the MC computation. Figure A.5 shows the *digital CTC* profile for the for genetic backgrounds.

The consistency of effective TRN subnetworks (TRN consistency) was calculated as the ratio of consistent links (i.e. the regulatory logic encoded on the links is consistent with the expression signs on the nodes; see Figure A.6) to overall effective links. Similar to the MC, this ratio was transformed into a z-score. Shuffling the expression signs of the effective nodes was used as a suitable null model. 5000 realizations of the null model were used for the z-score transformation.

Experimental setup

The *fis* and *hns* double mutants were generated by P1 transduction of mutant alleles from donor strains into the *E. coli* LZ41 and LZ54 strains used in previous study for investigation of the effects of single mutations [20]. The strains were grown in 2 × YT medium at 30°C. Total RNA isolated from exponentially growing LZ41Δ*fis*Δ*hns* and LZ54Δ*fis*Δ*hns* strains after brief (15 min) treatment by norfloxacin was subjected to DNA microarray-mediated transcription profiling using OciChip *E. coli* K12 V2 Arrays according to OciChipTM-Application Guide (http://www.ocimumbio.com) as described in Blot et al. [20].

In brief, for each comparison two biological replicates with two technical replicates were performed, resulting in a total of 8 hybridizations. Scanned array images were analyzed using the TM4 software package [102]. Spot intensities were quantified and the quality of each spot was verified by calculating a quality control (QC) score depending on signal-to-noise ratio for every channel and calculating p-values for each channel (as result of a t-test comparing the spot pixel set and surrounding background pixel set) using the TIGR Spotfinder software. Data was normalized by locally weighted linear regression [27]. A one-class t-test [91] was applied to obtain differentially expressed genes within each data set (significance level $\alpha < 0.05$).

Table A.1: Table of the MC values visualized in Figure 3.4

Network	WT	fis	hns	fis/hns
EcoCyc	3.37	0.31	-0.60	-0.65
EcoCyc*4	3.17	0.37	-0.53	-0.52
Intersec_EcoCyc_KEGG	3.03	0.26	0.54	-0.66
Intersec_EcoCyc_iAF1260man	2.89	0.57	-0.40	-0.13
iAF1260$^{*}_{man}$	2.75	1.47	0.85	-0.05
iAF1260_RichMedium	2.69	1.21	-1.11	-0.38
iAF1260$_{man}$	2.62	1.61	0.85	-0.065
EcoCyc**	2.55	0.22	0.33	0.65
KEGG*	2.57	0.76	1.59	1.32
iAF1260deg*	2.50	1.51	0.03	0.55
iAF1260deg	2.40	1.52	0.16	0.66
iAF1260man**	2.36	2.09	0.37	0.12
fully_coupled	2.22	0.14	1.70	1.14
KEGG	2.20	0.61	1.31	1.40
Intersec_iAF1260man_KEGG	1.80	1.43	-0.11	-0.46
Intersec_EcoCyc_iAF1260man_KEGG	1.76	0.68	-0.26	-0.48
KEGG**	1.58	-0.17	-0.46	1.048
fully_and_directionally_coupled	1.50	1.69	1.91	0.55
iAF1260$^{**}_{deg}$	1.37	1.62	0.14	-0.11
iAF1260**	1.20	1.17	0.42	1.17
directionally_coupled	1.06	0.39	0.03	1.51
iAF1260	0.19	1.17	0.23	1.40
iAF1260*	0.11	1.14	0.13	1.41

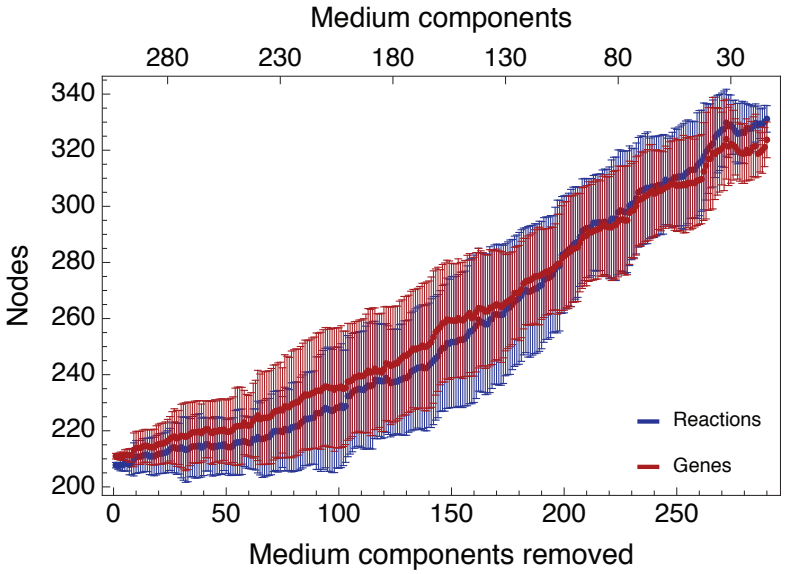


Figure A.1: The number of active reactions and genes in the effective networks increases when moving from rich to minimal media conditions. Only cytosolic reactions and genes were counted (i.e. transport and periplasmic reaction were excluded).

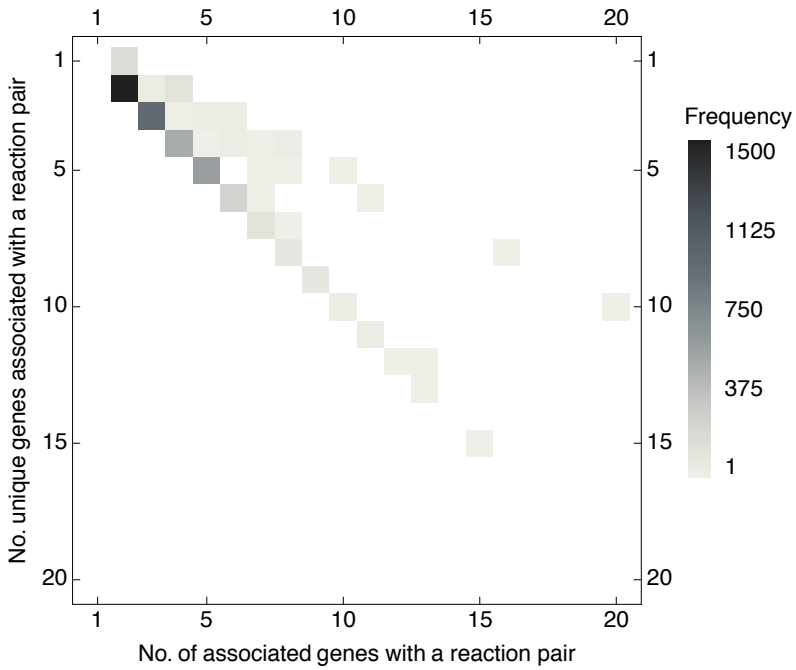


Figure A.2: The multiplicities of reaction-gene relations depicted as a colored grid.

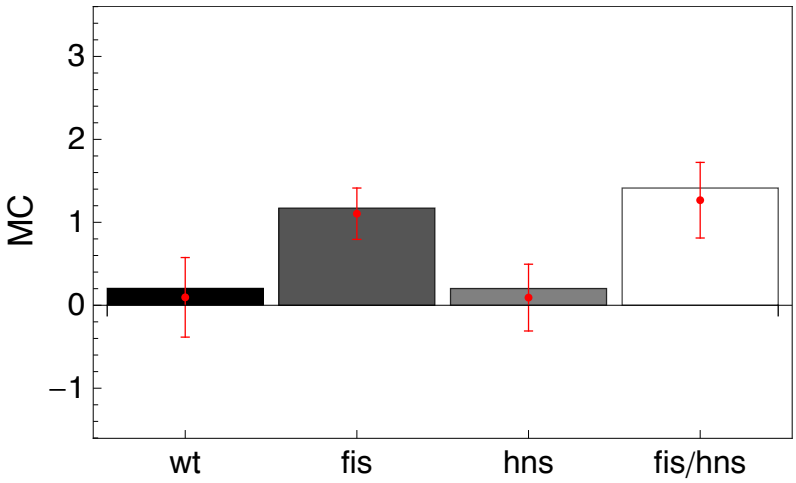


Figure A.3: Effect of currency metabolites as seen for the untreated *i*AF1260 network

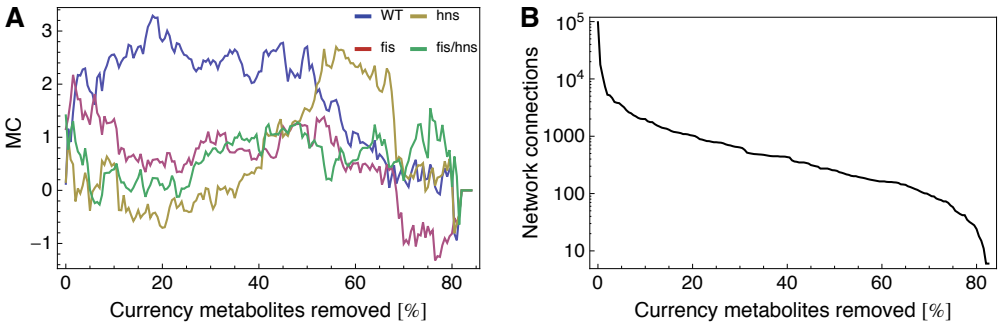


Figure A.4: (A) MC values plotted against the percentage of removed currency metabolites (as determined by the degree threshold method). (B) The network connectivity, i.e. number of connections in the network, plotted against the percentage of removed metabolites.

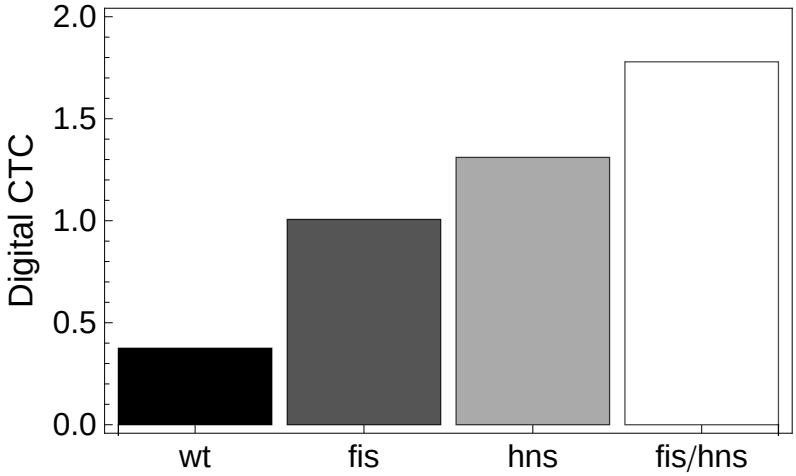


Figure A.5: Digital CTC (digital control) for the four genetic backgrounds.

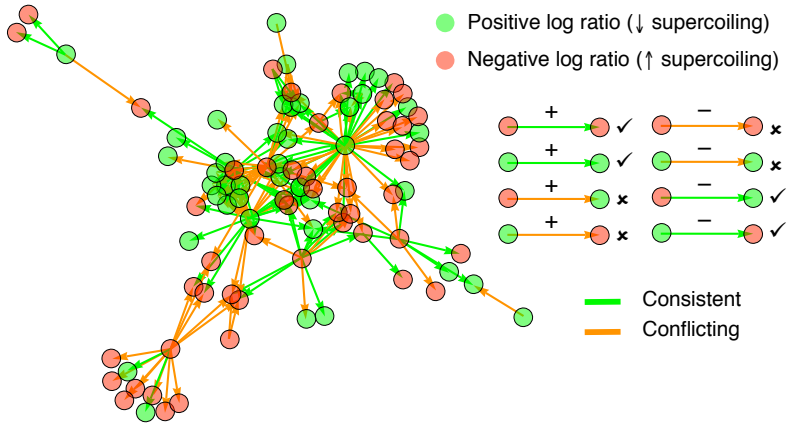


Figure A.6: The effective TRN (including only metabolic genes and their regulators) for the double mutant data (*fis*/*hns*). The scheme on the right-hand side explains the classification of consistent (checkmark; green link color) and inconsistent (x; orange link color) links.

Appendix B

Supplementary Information for "Multiple topological labels and medium-dependent essentiality in *Escherichia coli* metabolism"

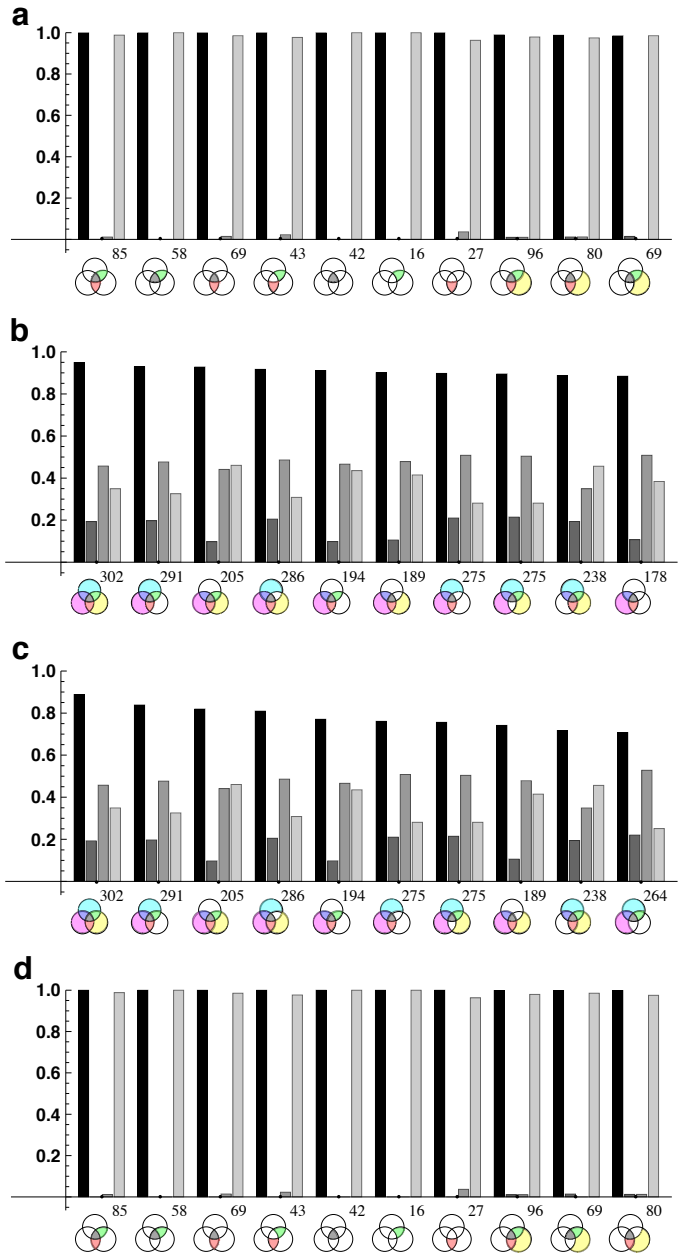


Figure B.1: Supplementary Figure 1. The other 10 highest performance scorers. The black bar represents each the (a) ppv, (b) npv, (c) sensitivity and (d) specificity of the intersection configurations. The other three colors represent the three essentiality classes with Class I being colored in dark gray and Class II and III in gray and light gray respectively.

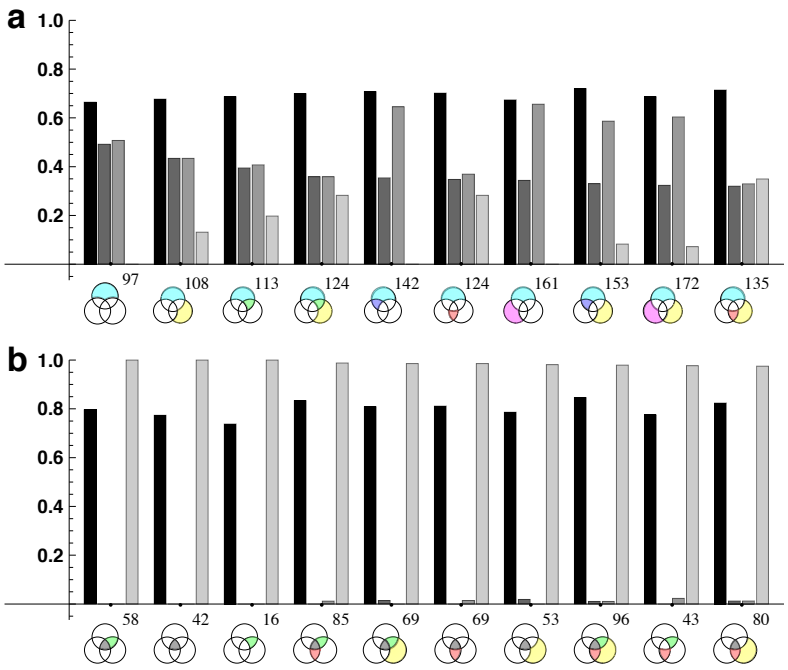


Figure B.2: Supplementary Figure 2. The top 10 (a) category I and (b) category II scorers. The black bar represents always the accuracy. The other three bar types represent the three essentiality classes with Class I being colored in dark gray and Class II and III in gray and light gray respectively.

Appendix C

Supplementary Information for "A network perspective on metabolic inconsistency"

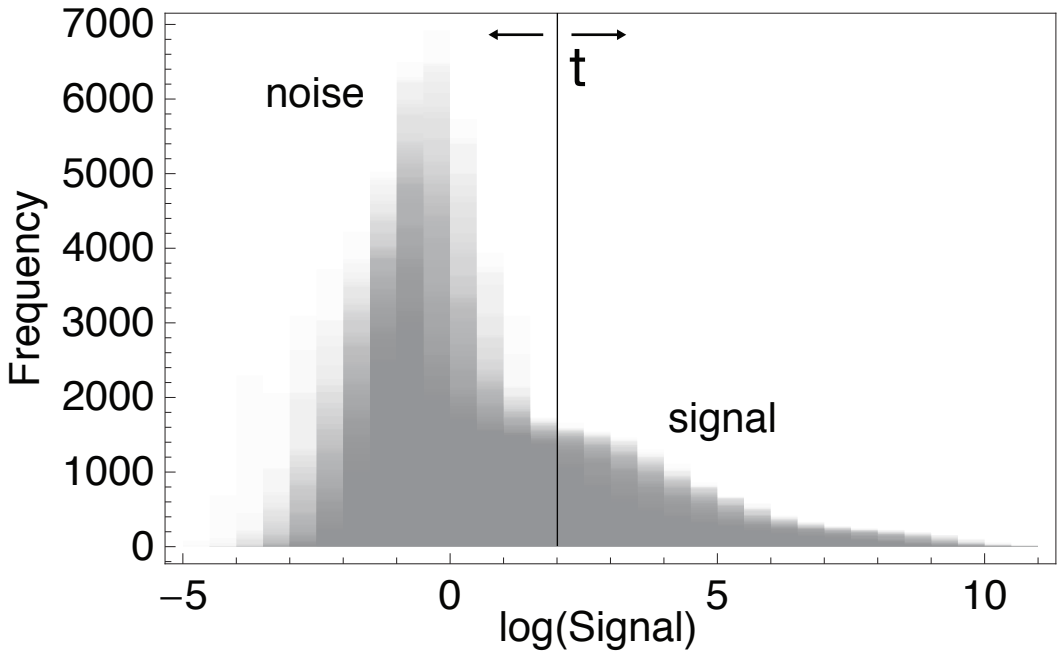


Figure C.1: Overlay of the logarithmized raw data of the 69 expression profiles. The bimodal distribution consisting of noise and signal becomes visible in this depiction.

Pathway maps

Pathways maps, referenced in Table 5.1 (SIaa, SIcarb, SIlip, SIvit), are provided electronically due to size limitations. The maps depict the usage patterns and inconsistency contributions for the overall contributions (page i), control (page ii), LIG (page iii), and HIG (page iv). The thickness and color of a reaction edge corresponds to the usage frequency and the contribution strength, respectively. The pathway maps have been obtained from the BIGG database [105].

SIaa (Amino acid metabolism) `http://goo.gl/exiq5`

SIcarb (Carbohydrate metabolism) `http://goo.gl/TaARR`

SIlip (Lipids metabolism) `http://goo.gl/hEc7b`

SIvit (Vita metabolism) `http://goo.gl/9qPHm`

Contributors

2-Oxoadipate producing pathways

2-Oxoadipate (2oxoadp) is one of the precursors for acetyl-CoA (see Figure C.3), which is heavily utilized in cholesterol biosynthesis. Only two paths lead to 2-oxoadipate (see Figure C.2 and C.3). The many missing GPR associations (*Invisible pathway*) on the path leading from lysine to 2-oxoadipate (Figure C.2), surely promote the usage of this specific pathway versus the alternative pathway leading from tryptophan to 2-oxoadipate (Figure C.3), explaining the high contribution strength of AATAi (see Figure 5.5 and Table 5.1). However, it is intriguing to see that allmost all subsequent steps from 2-oxoadipate to acetyl-CoA are expressed in the control and LIG.

PROD2: redox factor issue

Proline dehydrogenase (PROD2) is one of the major unspecific inconsistency contributors (see Table 5.1). Together with *pyrroline-5-carboxylate reductase* (P5CRx), it is involved in a cycle that interconverts NADH into FADH2 (see Figure C.4a), a necessary redox factor for cholesterol biosynthesis, which is the ultimate precursor for the steroid pathway and subsequent aldosterone production. P5CRxm is clearly expressed, whereas PROD2 is not expressed (see Figure C.4b). It is intriguing that the gene expression profiles are almost reversed in the mitochondrium (see Figure C.4b), where PROD2m is expressed (at least in the control) and P5CRxm is not expressed.

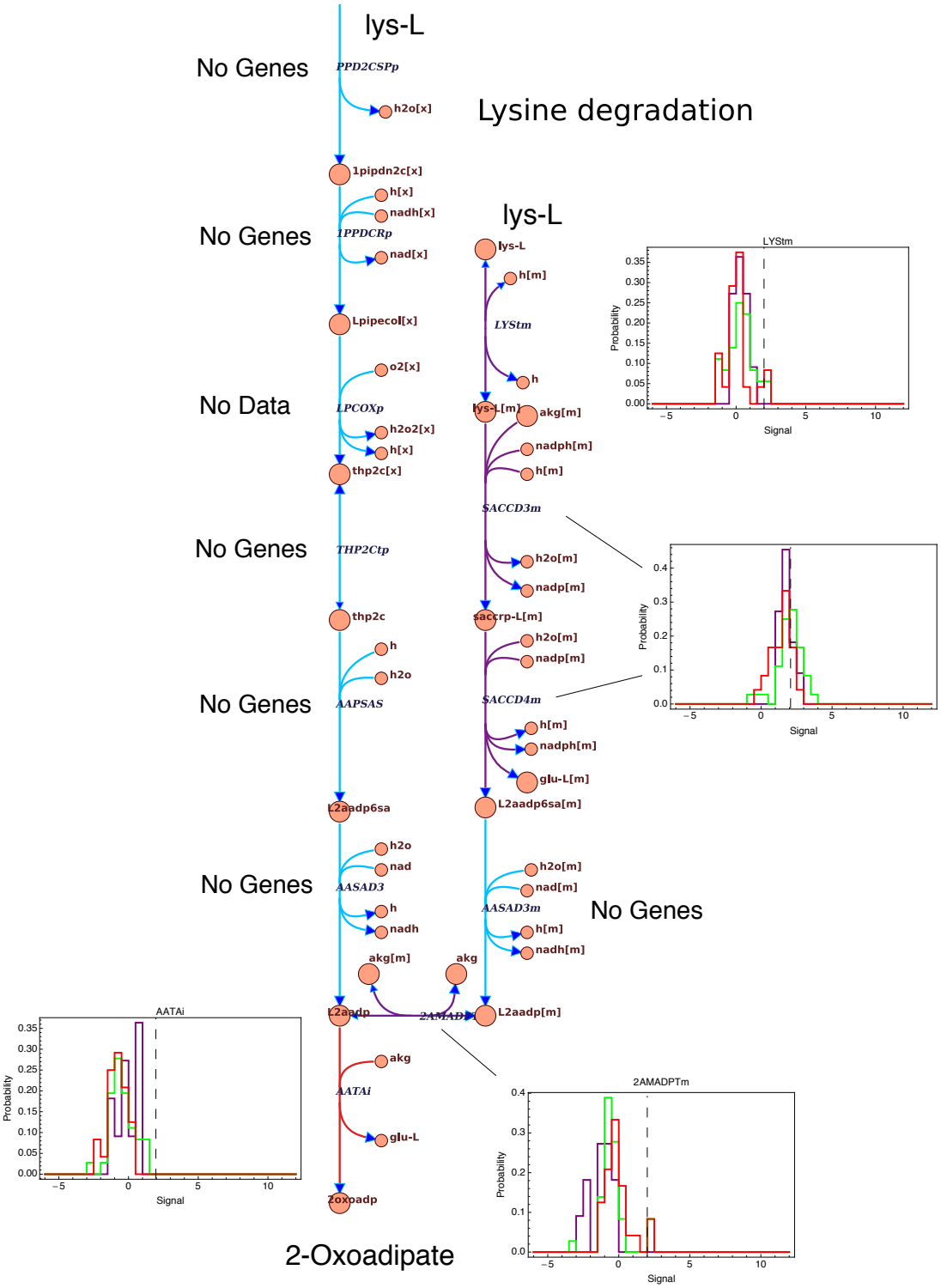


Figure C.2: 2-Oxoadipate production pathway starting from lysine and involving the unspecific contributor AATAi (see Table 5.1). Dashed lines indicate the threshold used for the GIMME computations.

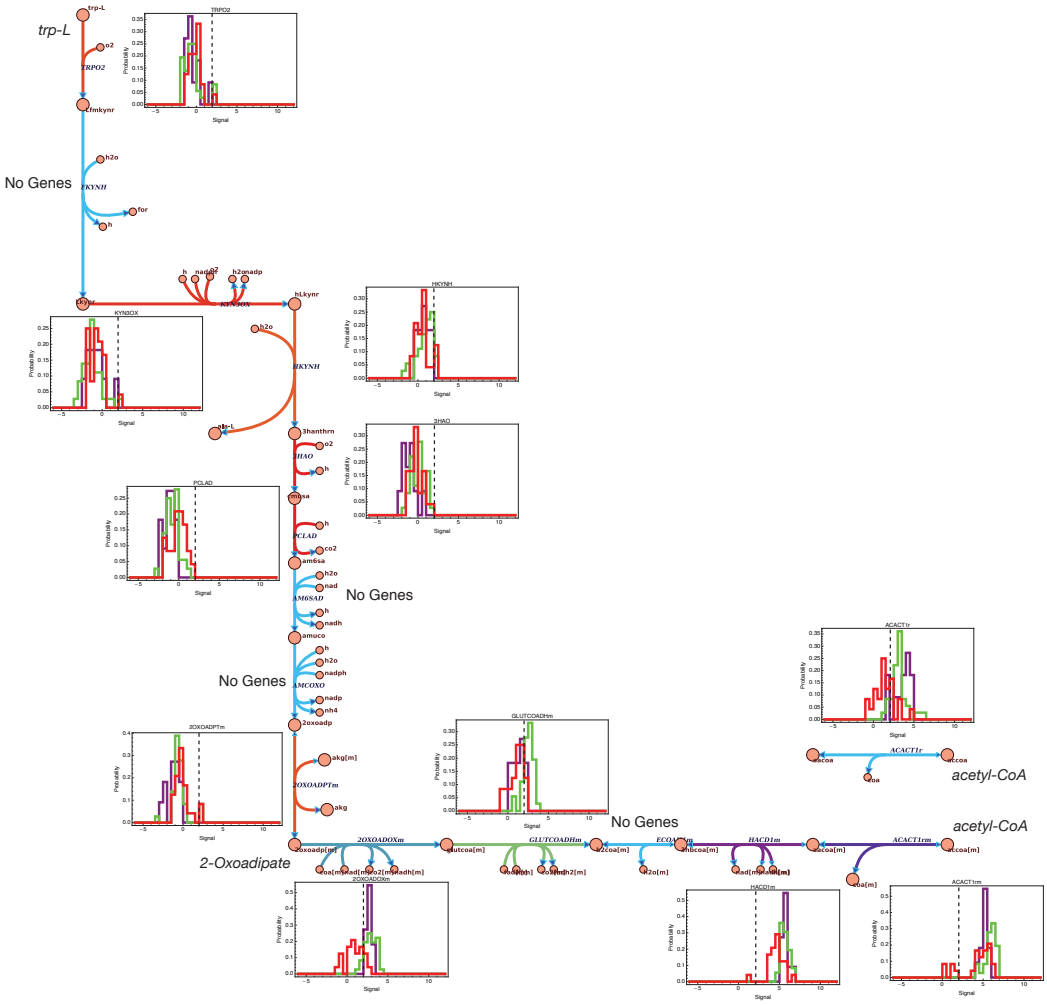
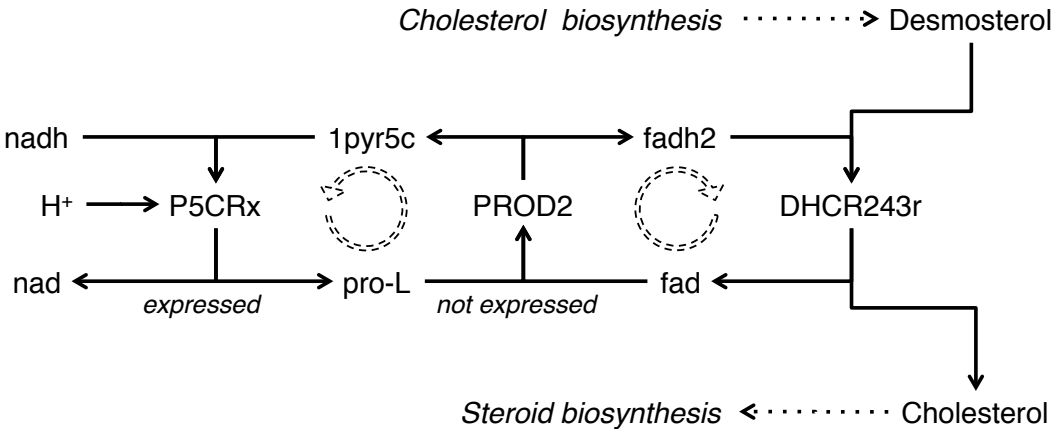
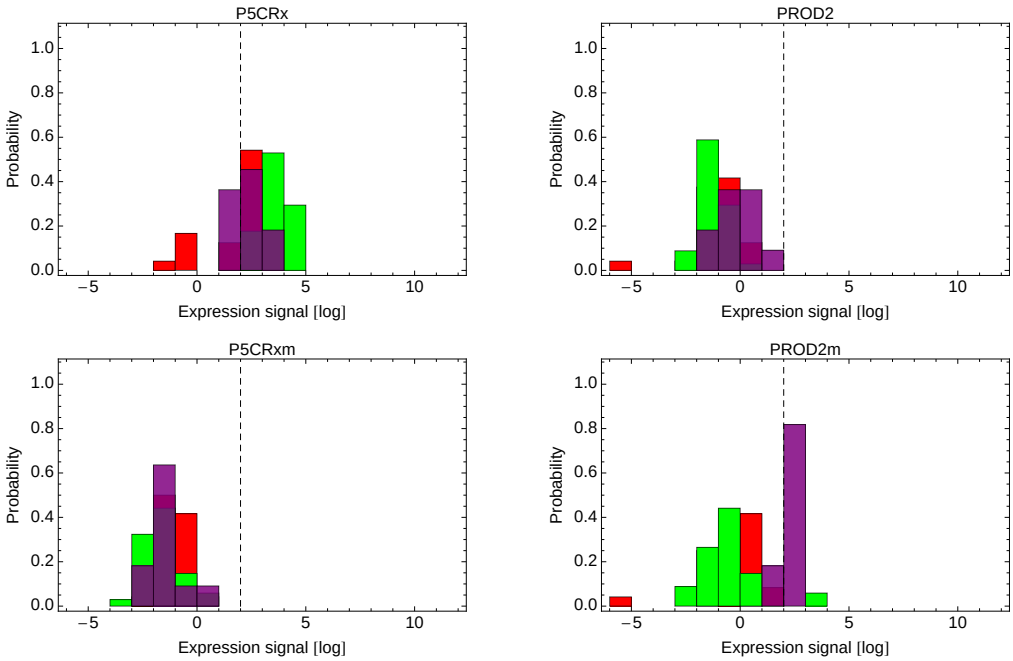


Figure C.3: 2-Oxoadipate production pathway starting from tryptophan and involving the unspecific contributor 2OXOADPTm (see Table 5.1) among other more specific contributions. Dashed lines indicate the threshold used for the GIMME computations.



(a)



(b)

Figure C.4: (a) The NADH to FADH2 interconverting cycle composed of *pyrroline-5-carboxylate reductase* (P5CRx, cytosolic) and *pyrroline-5-carboxylate reductase* (PROD2, cytosolic). (b) Distributions of expression values for the cytosolic (PROD2 and P5CRx) and mitochondrial (PROD2m and P5CRxm) versions of *pyrroline-5-carboxylate reductase* and *proline dehydrogenase*. The control is depicted in purple, the LIG and HIG in green and red, respectively, and the dashed lines indicate the threshold used for the GIMME computations.

Path from tyrosine to fumarate and acetoacetate

Tyrosine, which is not provided in the *in silico* medium (see Table C.1), and is build from phenylalanine under a strong contribution of PHETHPTOX2 (*phenylalanine 4-monooxygenase*), seems to be involved in another highly contributing pathway that leads from tyrosine to fumarate and acetoacetate (see Figure C.5). The entry and output point, TYRTA (*tyrosine transaminase*) and FUMAC (*fumarylacetoacetase*), of this chain of reactions seem to be expressed, whereas the intermediate steps, 34HPPOR (*4-hydroxyphenylpyruvate dioxygenase*), HGNTOR (*homogentisate 1,2-dioxygenase*), and MACACI (*maleylacetoacetate isomerase*), are all unspecific contributors enlisted in Table 5.1.

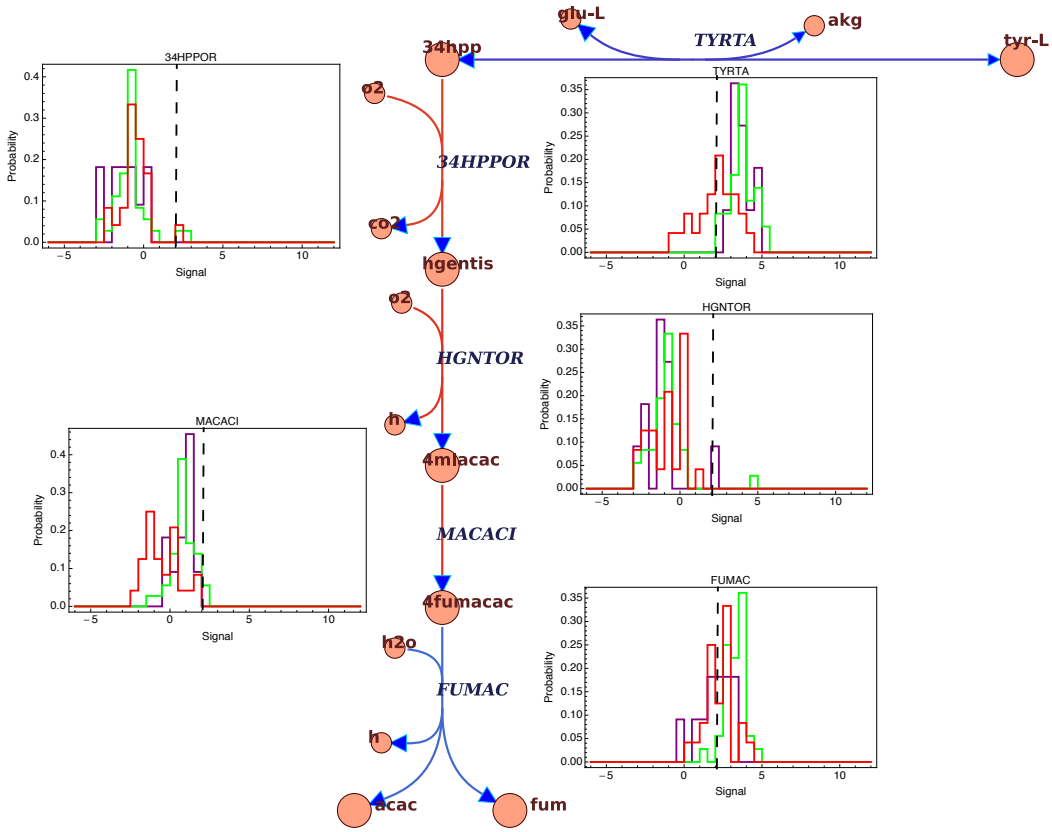


Figure C.5: Pathway leading from tyrosine to fumarate and acetoacetate. Dashed lines indicate the threshold used for the GIMME computations.

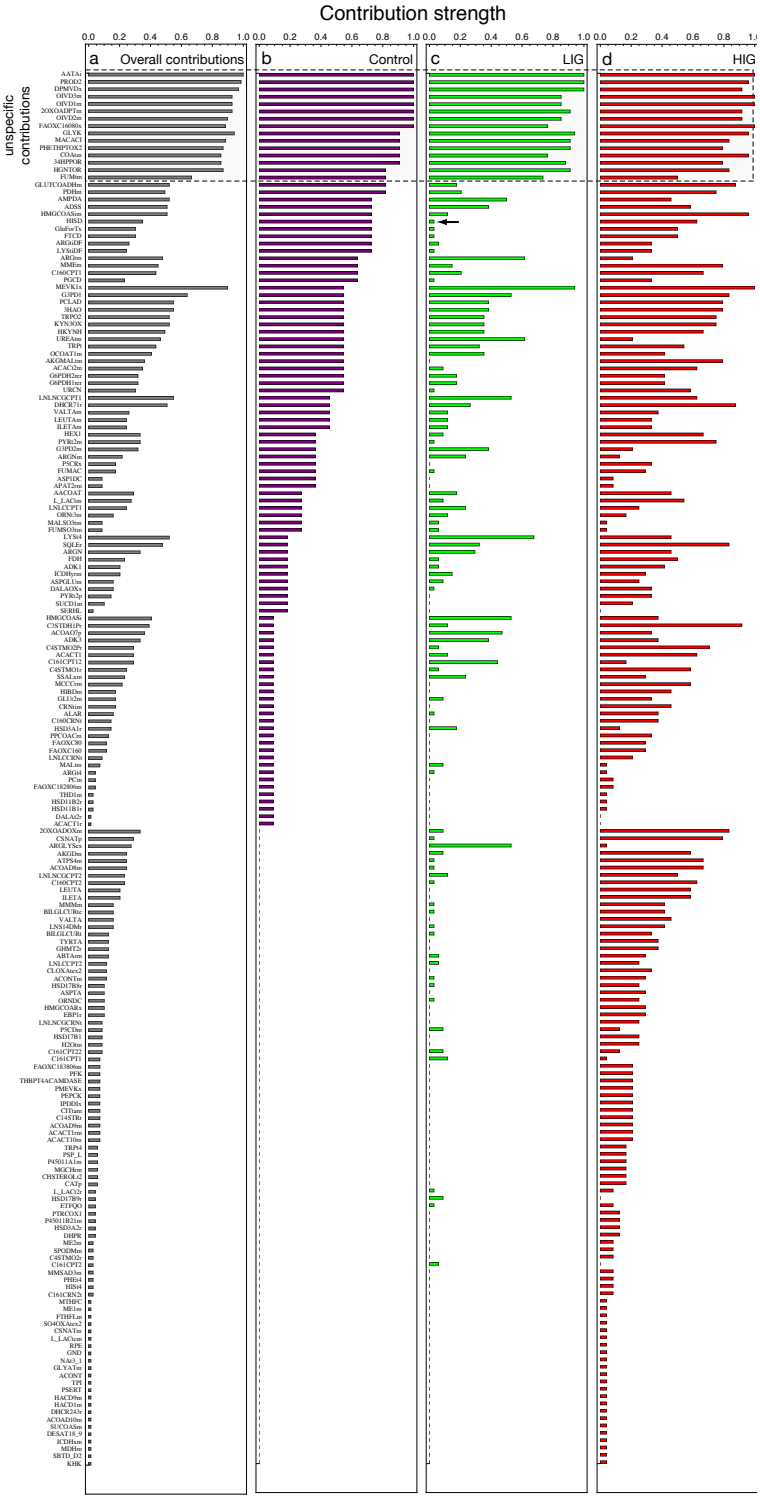


Figure C.6: Inconsistency contributions. (a) Overall contributions to the inconsistency for all adenoma and control samples, (b) the control group, and the adenoma tumor samples showing (c) lower and (d) higher inconsistencies in comparison to the control group. The contributions have been normalized by the sample size, respectively. Only a subset of all contributing reactions is shown due to space limitations.

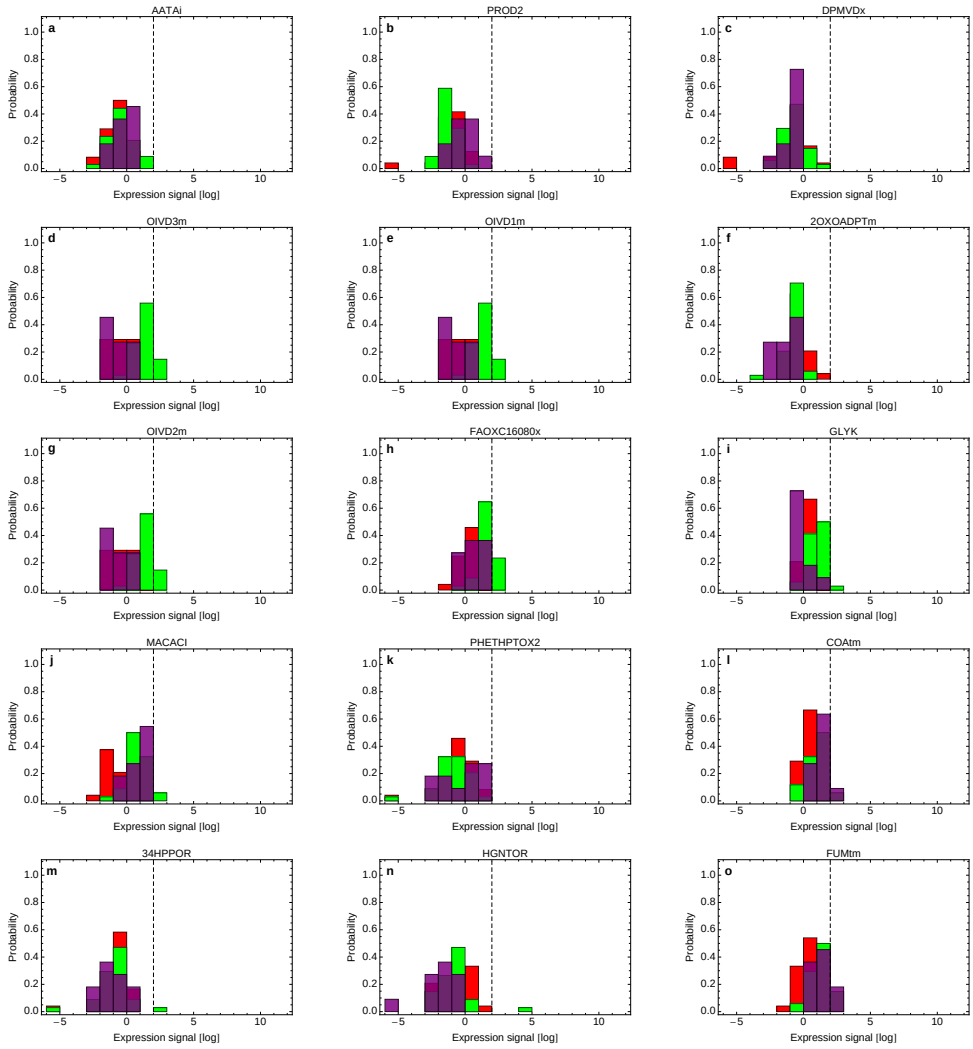


Figure C.7: Distributions of reactions expression values of the unspecific contributions in Table 5.1. The control is depicted in purple, the LIG and HIG in green and red, respectively, and the dashed line indicates the threshold used for the GIMME computations.

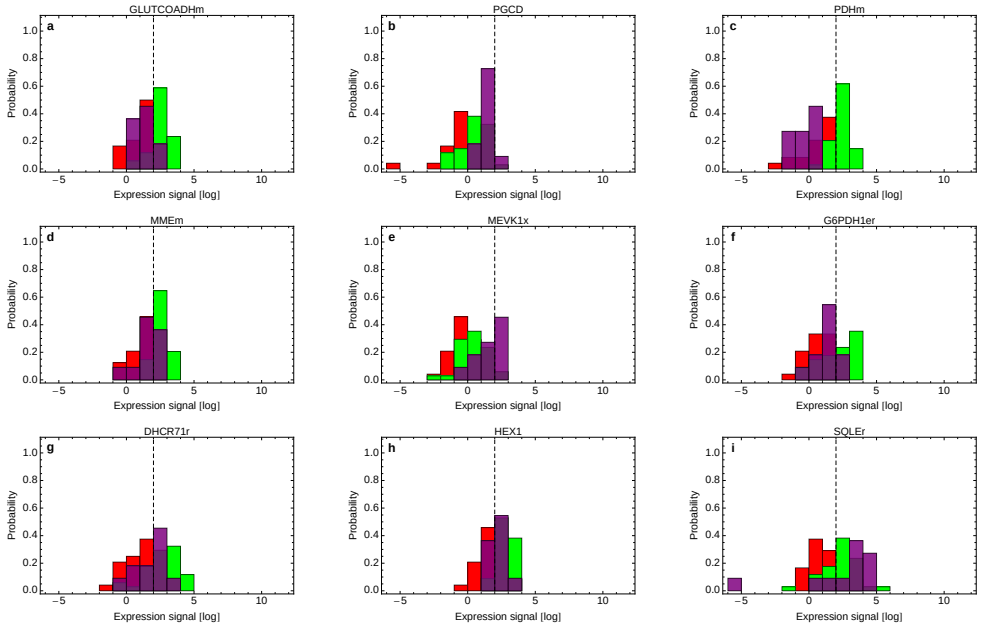


Figure C.8: Distributions of reactions expression values of the unspecific contributions in Table 5.1. The control is depicted in purple, the LIG and HIG in green and red, respectively, and the dashed line indicates the threshold used for the GIMME computations.

Table C.1: Medium condition from [113]. Exchange reactions not mentioned in the table default to a lower bound (uptake) of 0 and an upper bound (secretion) of 10000. DM reactions not mentioned in the table default to a lower(upper) bound of 0(10000).

Reaction name	Lower bound	Upper bound
DM_13-cis-oretn(n)	0	0
DM_13-cis-retn(n)	0	0
DM_avite1(c)	0	0
DM_avite2(c)	0	0
DM_bvite(c)	0	0
DM_yvite(c)	0	0
EX_arg-L(e)	-1	10000
EX_fe2(e)	0	0
EX_glc(e)	-1	10000
EX_glyc(e)	-1	10000
EX_h(e)	-1	10000
EX_h2o(e)	-1	10000
EX_hdca(e)	-1	10000
EX_his-L(e)	-1	10000
EX_ile-L(e)	-1	10000
EX_leu-L(e)	-1	10000
EX_lnlc(e)	-1	10000
EX_lys-L(e)	-1	10000
EX_met-L(e)	-1	10000
EX_o2(e)	-10000	10000
EX_phe-L(e)	-1	10000
EX_pi(e)	-1	10000
EX_so4(e)	-1	10000
EX_thr-L(e)	-1	10000
EX_trp-L(e)	-1	10000
EX_val-L(e)	-1	10000

References

[1] M. Akesson, J. Förster, and J. Nielsen. Integration of gene expression data into genome-scale metabolic models. *Metab Eng*, 6(4):285–93, 2004.

[2] R. Alberich, J. Miro-Julia, and F. Rossello. Marvel Universe looks almost like a real social network. *Arxiv preprint cond-mat/0202174*, 2002.

[3] R. Albert. Scale-free networks in cell biology. *J Cell Sci*, 118(Pt 21):4947–4957, 2005.

[4] R. Albert, H. Jeong, and A.-L. Barabasi. Internet: Diameter of the World-Wide Web. *Nature*, 401(6749):130–131, 1999.

[5] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

[6] T. E. Allen, N. D. Price, A. R. Joyce, and B. Ø. Palsson. Long-range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organization. *PLoS Comput Biol*, 2(1):e2, 2006.

[7] E. Almaas, B. Kovács, T. Vicsek, Z. N. Oltvai, and A.-L. Barabási. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature*, 427 (6977):839–843, 2004.

[8] E. Almaas, Z. N. Oltvai, and A.-L. Barabási. The activity reaction core and plasticity of metabolic networks. *PLoS Comput Biol*, 1(7):e68, 2005.

[9] C. Anderson. The end of theory: the data deluge makes the scientific method obsolete. *Wired Magazine*, July 2008. `http://goo.gl/aZsP`.

[10] M. Arita. The metabolic world of *Escherichia coli* is not small. *Proc Natl Acad Sci USA*, 101(6):1543–1547, 2004.

[11] H. Auner, M. Buckle, A. Deufel, T. Kutateladze, L. Lazarus, R. Mavathur, G. Muskhelishvili, I. Pemberton, R. Schneider, and A. Travers. Mechanism of transcriptional activation by FIS: role of core promoter structure and DNA topology. *J Mol Biol*, 331(2):331–44, 2003.

[12] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol*, 2: 2006.0008, 2006.

[13] B. J. Bachmann, B. K. Low, and A. L. Taylor. Recalibrated linkage map of *Escherichia coli* K-12. *Bacteriol Rev*, 40:116–167, 1976.

[14] V. L. Balke and J. D. Gralla. Changes in the linking number of supercoiled DNA accompany growth transitions in *Escherichia coli*. *J Bacteriol*, 169(10): 4499–506, 1987.

[15] J. Banga. Optimization in computational systems biology. *BMC Syst Biol*, 2(1): 47, 2008.

[16] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–12, 1999.

[17] A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–13, 2004.

[18] S. A. Becker and B. Ø. Palsson. Context-specific metabolic networks are consistent with experiments. *PLoS Comput Biol*, 4(5):e1000082, 2008.

[19] S. A. Becker, A. M. Feist, M. L. Mo, G. Hannum, B. Ø. Palsson, and M. J. Herrgard. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature Protocols*, 2(3):727–38, 2007.

[20] N. Blot, R. Mavathur, M. Geertz, A. Travers, and G. Muskhelishvili. Homeostatic regulation of supercoiling sensitivity coordinates transcription of the bacterial genome. *EMBO Rep*, 7(7):710–5, 2006.

[21] A. Bordbar, N. E. Lewis, J. Schellenberger, B. Ø. Palsson, and N. Jamshidi. Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. *Mol Syst Biol*, 6:422, 2010.

[22] S. Boulkroun, J.-F. G. Dzib, N. Sonnenschein, H. Leger, B. Samson-Couterie, A. Benecke, M.-T. Hütt, and M.-C. Zennaro. Using flux balance analysis to link gene expression with pathological profiles of a large cohort of human primary aldosteronism patients. *in preparation*, 2010.

[23] S. Boulkroun, B. Samson-Couterie, J.-F. G. Dzib, H. Lefebvre, E. Louiset, L. Amar, P.-F. Plouin, E. Lalli, X. Jeunemaitre, A. Benecke, T. Meatchi, and M.-C. Zennaro. Adrenal cortex remodeling and functional zona glomerulosa hyperplasia in primary aldosteronism. *Hypertension*, 56(5):885–92, 2010.

[24] A. P. Burgard, E. V. Nikolaev, C. H. Schilling, and C. D. Maranas. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res*, 14 (2):301–12, 2004.

[25] K. J. Cheung, V. Badarinarayana, D. W. Selinger, D. Janse, and G. M. Church. A microarray-based antibiotic screen identifies a regulatory role for supercoiling in the osmotic stress response of *Escherichia coli*. *Genome Res*, 13(2):206–15, 2003.

[26] M. Clarke. From the blogosphere. *Nature*, 454(x), July 2008. doi: dx.doi.org/10.1038/7201xc.

[27] W. Cleveland and S. Devlin. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

[28] C. Colijn, A. Brandes, J. Zucker, D. S. Lun, B. Weiner, M. R. Farhat, T.-Y. Cheng, D. B. Moody, M. Murray, and J. E. Galagan. Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. *PLoS Comput Biol*, 5(8):e1000489, 2009.

[29] M. W. Covert, E. Knight, J. L. Reed, M. J. Herrgard, and B. Ø. Palsson. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429(6987):92–96, 2004.

[30] P. Daran-Lapujade, S. Rossell, W. M. van Gulik, M. A. H. Luttik, M. J. L. de Groot, M. Slijper, A. J. R. Heck, J.-M. Daran, J. H. de Winde, H. V. Westerhoff, J. T. Pronk, and B. M. Bakker. The fluxes through glycolytic enzymes in *Saccharomyces cerevisiae* are predominantly regulated at posttranscriptional levels. *Proc Natl Acad Sci USA*, 104(40):15753–8, 2007.

[31] A. E. Darling, I. Miklós, and M. A. Ragan. Dynamics of genome rearrangement in bacterial populations. *PLoS Genet*, 4(7):e1000128, 2008.

[32] H. David, G. Hofmann, A. P. Oliveira, H. Jarmer, and J. Nielsen. Metabolic network driven analysis of genome-wide transcription data from *Aspergillus nidulans*. *Genome Biol*, 7(11):R108, 2006.

[33] Z. Dezso, Y. Nikolsky, E. Sviridov, W. Shi, T. Serebriyskaya, D. Dosymbekov, A. Bugrim, E. Rakhmatulin, R. J. Brennan, A. Guryanov, K. Li, J. Blake, R. R. Samaha, and T. Nikolskaya. A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol*, 6:49, 2008.

[34] N. Duarte, M. J. Herrgard, and B. Ø. Palsson. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res*, 14(7):1298–1309, 2004.

[35] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. Ø. Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci USA*, 104(6):1777–82, 2007.

[36] J. S. Edwards and B. Ø. Palsson. The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci USA*, 97(10):5528–5533, 2000.

[37] J. C. Engert. Unlimited hypothesis research. *Genome Res*, 10(3):271–2, 2000.

[38] G. Fang, E. P. C. Rocha, and A. Danchin. Persistence drives gene clustering in bacterial genomes. *BMC Genomics*, 9:4, 2008.

[39] A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. Ø. Palsson. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol*, 3:121, 2007.

[40] A. M. Feist, M. J. Herrgård, I. Thiele, J. L. Reed, and B. Ø. Palsson. Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol*, 7 (2):129–43, 2009.

[41] S. Gama-Castro, V. Jiménez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Peñaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muñiz-Rascado, I. Martínez-Flores, H. Salgado, C. Bonavides-Martinez, C. Abreu-Goodger, C. Rodríguez-Penagos, J. Miranda-Ríos, E. Morett, E. Merino, A. M. Huerta, L. Treviño-Quintanilla, and J. Collado-Vides. RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res*, 36 (Database issue):D120–4, 2008.

[42] S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muñiz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. García-Sotelo,

A. López-Fuentes, L. Porrón-Sotelo, S. Alquicira-Hernández, A. Medina-Rivera, I. Martínez-Flores, K. Alquicira-Hernández, R. Martínez-Adame, C. Bonavides-Martínez, J. Miranda-Ríos, A. M. Huerta, A. Mendoza-Vargas, L. Collado-Torres, B. Taboada, L. Vega-Alvarado, M. Olvera, L. Olvera, R. Grande, E. Morett, and J. Collado-Vides. RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res*, 2010. doi: 10.1093/nar/gkq1110.

[43] S. Gerdes, M. D. Scholle, J. W. Campbell, G. Balázsi, E. Ravasz, M. D. Daugherty, A. L. Somera, N. C. Kyrpides, I. Anderson, M. S. Gelfand, A. Bhattacharya, V. Kapatral, M. D'Souza, M. V. Baev, Y. Grechkin, F. Mseeh, M. Fonstein, R. Overbeek, A.-L. Barabási, Z. N. Oltvai, and A. L. Osterman. Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J Bacteriol*, 185(19):5673–84, 2003.

[44] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási. The human disease network. *Proc Natl Acad Sci USA*, 104(21):8685–90, 2007.

[45] L. Goodman. Hypothesis-limited research. *Genome Res*, 9(8):673–4, 1999.

[46] J. Gowrishankar and R. Harinarayanan. Why is transcription coupled to translation in bacteria? *Mol Microbiol*, 54(3):598–603, 2004.

[47] R. Guimerà and L. N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

[48] R. Guimerà, S. Mossa, A. Turtschi, and L. A. N. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proc Natl Acad Sci USA*, 102(22):7794–9, 2005.

[49] R. Guimera, B. Uzzi, J. Spiro, and L. A. N. Amaral. Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science*, 308(5722):697–702, 2005.

[50] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral. Classes of complex networks defined by role-to-role connectivity profiles. *Nature Physics*, 3(1):63–69, 2007.

[51] R. M. Gutiérrez-Ríos, D. A. Rosenblueth, J. A. Loza, A. M. Huerta, J. D. Glasner, F. R. Blattner, and J. Collado-Vides. Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res*, 13(11):2435–43, 2003.

[52] T. Handorf, O. Ebenhoh, and R. Heinrich. Expanding metabolic networks: scopes of compounds, robustness, and evolution. *J Mol Evol*, 61(4):498–512, 2005.

[53] C. D. Hardy and N. R. Cozzarelli. A genetic selection for supercoiling mutants of *Escherichia coli* reveals proteins implicated in chromosome structure. *Mol Microbiol*, 57(6):1636–52, 2005.

[54] R. Harrison, B. Papp, C. Pál, S. Oliver, and D. Delneri. Plasticity of genetic interactions in metabolic networks of yeast. *Proc Natl Acad Sci USA*, 104(7): 2307–2312, 2007.

[55] G. W. Hatfield and C. J. Benham. DNA topology-mediated control of global gene expression in *Escherichia coli*. *Annu Rev Genet*, 36:175–203, 2002.

[56] R. Hermsen, P. R. ten Wolde, and S. Teichmann. Chance and necessity in chromosomal gene distributions. *Trends Genet*, 24(5):216–9, 2008.

[57] M. J. Herrgård, M. W. Covert, and B. Ø. Palsson. Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res*, 13 (11):2423–34, 2003.

[58] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. L. Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang, and S. Forum. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–31, 2003.

[59] M. Huss and P. Holme. Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks. *IET systems biology*, 1(5):280–5, 2007.

[60] M. T. Hütt and A. Lesne. Interplay between topology and dynamics in excitation patterns on hierarchical graphs. *Front Neuroinformatics*, 3:28, 2009.

[61] T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929–34, 2001.

[62] J. Illian, A. Penttinen, and H. Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley-Interscience, Chichester, 2008.

[63] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3:318–56, 1961.

[64] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[65] K. S. Jeong, J. Ahn, and A. B. Khodursky. Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. *Genome Biol*, 5(11):R86, 2004.

[66] A. R. Joyce and B. Ø. Palsson. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol*, 7(3):198–210, 2006.

[67] A. R. Joyce and B. Ø. Palsson. Predicting gene essentiality using genome-scale in silico models. *Methods Mol Biol*, 416:433–57, 2008.

[68] J. Jurka and M. A. Savageau. Gene density over the chromosome of *Escherichia coli*: frequency distribution, spatial clustering, and symmetry. *J Bacteriol*, 163 (2):806–11, 1985.

[69] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue):D480–4, 2008.

[70] P. D. Karp, I. M. Keseler, A. Shearer, M. Latendresse, M. Krummenacker, S. M. Paley, I. Paulsen, J. Collado-Vides, S. Gama-Castro, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Peñaloza-Spínola, C. Bonavides-Martinez, and J. Ingraham. Multidimensional annotation of the *Escherichia coli* K-12 genome. *Nucleic Acids Res*, 35(22):7577–90, 2007.

[71] K. J. Kauffman, P. Prakash, and J. S. Edwards. Advances in flux balance analysis. *Curr Opin Biotechnol*, 14(5):491–6, 2003.

[72] F. Képès. Periodic transcriptional organization of the E.coli genome. *J Mol Biol*, 340(5):957–64, 2004.

[73] P. Kharchenko, G. M. Church, and D. Vitkup. Expression dynamics of a cellular metabolic network. *Molecular Systems Biology*, 1:2005.0016, 2005.

[74] P.-J. Kim, D.-Y. Lee, T. Y. Kim, K. H. Lee, H. Jeong, S. Y. Lee, and S. Park. Metabolite essentiality elucidates robustness of *Escherichia coli* metabolism. *Proc Natl Acad Sci USA*, 104(34):13638–42, 2007.

[75] A. G. Ladurner. Chromatin places metabolism center stage. *Cell*, 138(1):18–20, 2009.

[76] 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(7006):308–12, 2004.

[77] H. Ma and A. Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19 (2):270–277, 2003.

[78] H.-W. Ma, B. Kumar, U. Ditges, F. Gunzer, J. Buer, and A.-P. Zeng. An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs. *Nucleic Acids Res*, 32(22):6643–9, 2004.

[79] R. Mahadevan and B. Ø. Palsson. Properties of metabolic networks: structure versus function. *Biophys J*, 88(1):07–09, 2005.

[80] R. Mahadevan, J. S. Edwards, and F. Doyle. Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophys J*, 83(3):1331–1340, 2002.

[81] C. Marr, M. Geertz, M.-T. Hütt, and G. Muskhelishvili. Dissecting the logical types of network control in gene expression profiles. *BMC Syst Biol*, 2:18, 2008.

[82] D. A. D. Martelaere and A. P. V. Gool. The density distribution of gene loci over the genetic map of *Escherichia coli*: its structural, functional and evolutionary implications. *J Mol Evol*, 17(6):354–60, 1981.

[83] R. Montañez, M. A. Medina, R. V. Solé, and C. Rodríguez-Caso. When metabolism meets topology: Reconciling metabolite and reaction networks. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 32(3): 246–56, 2010.

[84] G. Muskhelishvili and A. Travers. Intrinsic in vivo modulators: negative supercoiling and the constituents of the bacterial nucleoid. In H. Buc and T. Strick, editors, *RNA polymerases as molecular motors*, chapter 3, pages 69–95. RSC Publishing, Cambridge, UK, 2009.

[85] G. Muskhelishvili, P. Sobetzko, M. Geertz, and M. Berger. General organisational principles of the transcriptional regulation system: a tree or a circle? *Mol Biosyst*, pages 1–20, 2010.

[86] R. A. Notebaart, B. Teusink, R. J. Siezen, and B. Papp. Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLoS Comput Biol*, 4(1):e26, 2008.

[87] M. L. Opel, K. A. Aeling, W. M. Holmes, R. C. Johnson, C. J. Benham, and G. W. Hatfield. Activation of transcription initiation from a stable RNA promoter by a Fis protein-mediated DNA structural transmission mechanism. *Mol Microbiol*, 53(2):665–74, 2004.

[88] C. Pal, B. Papp, and M. Lercher. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet*, 37(12):1372–1375, 2005.

[89] G. Palla, A.-L. Barabási, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–7, 2007.

[90] B. Ø. Palsson. *Systems biology: properties of reconstructed networks*. Cambridge University Press New York, NY, USA, 2006. ISBN 0521859034.

[91] W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18 (4):546–54, 2002.

[92] J. A. Papin, N. D. Price, S. Wiback, D. Fell, and B. Ø. Palsson. Metabolic pathways in the post-genome era. *Trends Biochem Sci*, 28(5):250–258, 2003.

[93] B. Papp, C. Pál, and L. Hurst. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature*, 429(6992):661–664, 2004.

[94] B. J. Peter, J. Arsuaga, A. M. Breier, A. B. Khodursky, P. O. Brown, and N. R. Cozzarelli. Genomic transcriptional response to loss of chromosomal supercoiling in *Escherichia coli*. *Genome Biol*, 5(11):R87, 2004.

[95] N. D. Price, J. L. Reed, and B. Ø. Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol*, 2(11): 886–897, 2004.

[96] E. Ravasz, A. L. Somera, D. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586): 1551–1555, 2002.

[97] J. L. Reed, T. D. Vo, C. H. Schilling, and B. Ø. Palsson. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol*, 4(9): R54, 2003.

[98] E. P. C. Rocha. The organization of the bacterial genome. *Annu Rev Genet*, 42: 211–33, 2008.

[99] M. Rochman, M. Aviv, G. Glaser, and G. Muskhelishvili. Promoter protection by a transcription factor acting as a local topological homeostat. *EMBO Rep*, 3 (4):355–60, 2002.

[100] P. Romero and P. D. Karp. Nutrient-related analysis of pathway/genome databases. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pages 471–82, 2001.

[101] S. Rossell, C. C. van der Weijden, A. Lindenbergh, A. van Tuijl, C. Francke, B. M. Bakker, and H. V. Westerhoff. Unraveling the complexity of flux regulation: a new method demonstrated for nutrient starvation in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA*, 103(7):2166–71, 2006.

[102] A. I. Saeed, V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, A. Sturn, M. Snuffin, A. Rezantsev, D. Popov, A. Ryltsov, E. Kostukovich, I. Borisovsky, Z. Liu, A. Vinsavich, V. Trush, and J. Quackenbush. TM4: a free, open-source system for microarray data management and analysis. *BioTechniques*, 34(2):374–8, 2003.

[103] A. Samal and S. Jain. The regulatory network of *E. coli* metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response. *BMC Syst Biol*, 2:21, 2008.

[104] A. Samal, S. Singh, V. Giri, S. Krishna, N. Raghuram, and S. Jain. Low degree metabolites explain essential reactions and enhance modularity in biological networks. *BMC Bioinformatics*, 7:118, 2006.

[105] J. Schellenberger, J. O. Park, T. M. Conrad, and B. Ø. Palsson. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, 11(1):213, 2010.

[106] R. Schneider, A. Travers, and G. Muskhelishvili. The expression of the *Escherichia coli* fis gene is strongly dependent on the superhelical density of DNA. *Mol Microbiol*, 38(1):167–75, 2000.

[107] J.-M. Schwartz and M. Kanehisa. Quantitative elementary mode analysis of metabolic pathways: the example of yeast glycolysis. *BMC Bioinformatics*, 7(1): 186, 2006.

[108] D. Segrè, D. Vitkup, and G. Church. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci USA*, 99(23):15112–15117, 2002.

[109] S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet*, 31(1):64–68, 2002.

[110] T. Shlomi, O. Berkman, and E. Ruppin. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci USA*, 102(21):7695–7700, 2005.

[111] T. Shlomi, Y. Eisenberg, R. Sharan, and E. Ruppin. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol Syst Biol*, 3:101, 2007.

[112] T. Shlomi, M. Cabili, M. Herrgård, B. Ø. Palsson, and E. Ruppin. Network-based prediction of human tissue-specific metabolism. *Nat Biotechnol*, 2008.

[113] M. I. Sigurdsson, N. Jamshidi, J. J. Jonsson, and B. O. Palsson. Genome-scale network analysis of imprinted human metabolic genes. *Epigenetics : official journal of the DNA Methylation Society*, 4(1):43–6, 2009.

[114] J. L. Snoep, C. C. van der Weijden, H. W. Andersen, H. V. Westerhoff, and P. R. Jensen. DNA supercoiling in *Escherichia coli* is under tight and subtle homeostatic control, involving gene-expression and metabolic regulation of both topoisomerase I and DNA gyrase. *Eur J Biochem*, 269(6):1662–9, 2002.

[115] N. Sonnenschein, M.-T. Hütt, H. Stoyan, and D. Stoyan. Ranges of control in the transcriptional regulation of *Escherichia coli*. *BMC Syst Biol*, 3(1):119, 2009.

[116] N. Sonnenschein, A. Benecke, A. Lesne, and M.-T. Hütt. A network perspective on metabolic inconsistency. *in preparation*, 2010.

[117] N. Sonnenschein, M. Geertz, G. Muskhelishvili, and M.-T. Hütt. Analog regulation of metabolic demand. *submitted*, 2010.

[118] N. Sonnenschein, C. Marr, and M.-T. Hütt. Multiple topological labels and medium-dependent essentiality in *Escherichia coli* metabolism. *in preparation*, 2010.

[119] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–193, 2002.

[120] S. H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–76, 2001.

[121] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–8, 2010.

[122] I. Thiele, N. D. Price, T. D. Vo, and B. Ø. Palsson. Candidate metabolic network states in human mitochondria. Impact of diabetes, ischemia, and diet. *J Biol Chem*, 280(12):11683–95, 2005.

[123] A. Travers and G. Muskhelishvili. DNA microloops and microdomains: a general mechanism for transcription activation by torsional transmission. *J Mol Biol*, 279(5):1027–43, 1998.

[124] A. Travers and G. Muskhelishvili. DNA supercoiling - a global transcriptional regulator for enterobacterial growth? *Nat Rev Microbiol*, 3(2):157–69, 2005.

[125] A. Travers and G. Muskhelishvili. Bacterial chromatin. *Curr Opin Genet Dev*, 15(5):507–14, 2005.

[126] R. Urbanczik. SNA–a toolbox for the stoichiometric analysis of metabolic networks. *BMC Bioinformatics*, 7:129, 2006.

[127] M. van Workum, S. J. van Dooren, N. Oldenburg, D. Molenaar, P. R. Jensen, J. L. Snoep, and H. V. Westerhoff. DNA supercoiling depends on the phosphorylation potential in *Escherichia coli*. *Mol Microbiol*, 20(2):351–60, 1996.

[128] A. Varma and B. Ø. Palsson. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl Environ Microbiol*, 60(10):3724–3731, 1994.

[129] L. von Bertalanffy. *General system theory: foundations, development, applications*. George Braziller, New York, NY, USA, 1973.

[130] T. Vora, A. K. Hottes, and S. Tavazoie. Protein occupancy landscape of a bacterial genome. *Mol Cell*, 35(2):247–53, 2009.

[131] A. Wagner and D. A. Fell. The small world inside large metabolic networks. *Proc Biol Sci*, 268(1478):1803–10, 2001.

[132] P. B. Warren and P. R. ten Wolde. Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of *Escherichia coli*. *J Mol Biol*, 342(5):1379–90, 2004.

[133] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–2, 1998.

[134] S. J. Wiback, R. Mahadevan, and B. Ø. Palsson. Using metabolic flux data to further constrain the metabolic solution space and predict internal flux patterns: the *Escherichia coli spectrum*. *Biotechnol Bioeng*, 86(3):317–31, 2004.

[135] M. A. Wright, P. Kharchenko, G. M. Church, and D. Segrè. Chromosomal periodicity of evolutionarily conserved gene pairs. *Proc Natl Acad Sci USA*, 104 (25):10559–64, 2007.

[136] Z. Wunderlich and L. A. Mirny. Using the topology of metabolic networks to predict viability of mutant strains. *Biophys J*, 91(6):2304–11, 2006.

[137] S.-H. Yook, H. Jeong, and A.-L. Barabási. Modeling the Internet's large-scale topology. *Proc Natl Acad Sci USA*, 99(21):13382–6, 2002.

[138] H. Yu and M. Gerstein. Genomic analysis of the hierarchical structure of regulatory networks. *Proc Natl Acad Sci USA*, 103(40):14724–31, 2006.

[139] A. Zaslaver, A. Mayo, R. Rosenberg, P. Bashkin, H. Sberro, M. Tsalyuk, M. Surette, and U. Alon. Just-in-time transcription program in metabolic pathways. *Nat Genet*, 36(5):486–491, 2004.

[140] E. L. Zechiedrich, A. B. Khodursky, and N. R. Cozzarelli. Topoisomerase IV, not gyrase, decatenates products of site-specific recombination in *Escherichia coli*. *Genes Dev*, 11(19):2580–92, 1997.

Acknowledgments

This thesis would not have been possible without my close collaborators, Georgi Muskhelishvili, Arndt Benecke, Annick Lesne, Dietrich Stoyan, Marcel Geertz, and Carsten Marr. I thank you for your advice, criticism, and cooperation.

Kudos to the following people, for challenging me in endless discussions, proofreading my poor man's english, and general support: Eva Käppel (soon to be Sonnenschein), Moritz Beber, Daniel Geberth, James Smith, and Miriam Grace. Furthermore, I am indebted to Jacobs University and Volkswagen Stiftung for funding my research.

Last, but most importantly, I owe my mentor Marc Hütt for saving me from pipetting hell and giving me the opportunity to work as a computational biologist. Though, you certainly will not miss my annoyances and occasional palaver, be sure that I will miss your trust, forgiveness, respect, patience, and integrity.

I hereby confirm that the thesis presented here has not been submitted at another university for the conferral of a degree, and was written by me, Nikolaus Sonnenschein, using only the cited sources.

Bremen, January 14, 2011