A Systems Biology Approach to the Evolution of Codon Use Pattern

Ines Thiele$^{1,2,*}$, Ronan M.T. Fleming$^{1,3}$, Richard Que$^4$, Aarash Bordbar$^4$, Bernhard O. Palsson$^4$

August 31, 2011

1. Center for Systems Biology, University of Iceland, Reykjavik, Iceland
2. Faculty of Industrial Engineering, Mechanical Engineering & Computer Science, University of Iceland, Reykjavik, Iceland
3. Science Institute, University of Iceland, Reykjavik, Iceland
4. Department of Bioengineering, University of California, San Diego, CA, USA

* Corresponding Author: Ines Thiele, ithiele@hi.is.

Keywords: metabolism, codon usage, macromolecular synthesis, evolution, constraint-based modeling, multi-scale modeling

Running Title: Systems Biology of Codon Usage

character count (including spaces): 28,524
Abstract

The genetic code is redundant as amino acids are encoded by synonymous codons that are unequally used. This codon usage bias (CUB) affects gene expression and cellular functions yet the underlying mechanisms have not been elucidated. We used a sequence-specific, stoichiometric model of metabolism and macromolecular synthesis for *Escherichia coli* K12 MG1655 to test the effect of randomly changed CUB on growth maximization under various environmental conditions. Amongst CUB mutant strains, we identified reduced growth phenotypes, which were caused by tRNA supply shortage. We propose, supported by computations and bibliomic data, that expansion of tRNA gene content or tRNA reading is a mechanism to respond to changes in CUB. Our systems biology modelling framework suggests that in order to maximize growth and to adapt to new environmental niches, CUB and tRNA content must co-evolve and provides further evidence for the mutation-selection-drift balance theory of CUB.
Introduction

The degeneracy of the genetic code implies that one or more cognate transfer RNA (tRNA) species can recognize the same codon (a triplet of nucleotides using a four letter code) on a messenger RNA (mRNA), while a tRNA species can also read two or more synonymous codons. There is a unique set of codons and tRNA species per amino acid. The number of amino acids and codons is fixed to 20 and 64, respectively, but the number of tRNA genes varies widely (29-126) even between closely related organisms. Fast growing bacteria contain a higher number of tRNA genes for a smaller set of the possible anticodons (corresponding triplets on the tRNA species) (Rocha, 2004). At the same time, the frequency of synonymous codon use differs between organisms, within genomes, and along genes, a phenomenon known as CUB.

Two non-exclusive hypotheses have been proposed to explain co-evolution of CUB and tRNA content (Plotkin and Kudla, 2011; Hershberg and Petrov, 2008; Bulmer, 1991): i) the mutational (neutral) hypothesis proposes that mutational processes without any associated loss or gain of function occur (e.g., through changes in cellular nucleotide content (Hershberg et al, 2009) leading to changes in CUB); and ii) the natural selection hypothesis suggests that synonymous mutations affect the fitness of the organism and manifest in CUB across the genome or genes (Plotkin and Kudla, 2011; Elf et al, 2003). Current empirical and experimental evidence provides support for both hypotheses, also known as mutation-selection-drift balance theory of CUB (Bulmer, 1991). No comprehensive conceptual framework exists to investigate the link between CUB and tRNA content and its effect on protein synthesis, growth phenotype, and possible growth environments.

Cell-scale modeling is one of the great goals of computational biology. In fact, in 2002 an international *Escherichia coli* alliance was formed with the aim to generate data and tools necessary to formulate a whole cell computer representation of this bacteria (Holden, 2002). In the constraint-based reconstruction and analysis (COBRA) approach, biochemical reactions networks are reconstructed based on an organism’s genome and physico-chemical principles (Thiele and Palsson, 2010; Palsson, 2006). Bottom-up network reconstructions have been developed for metabolism (Feist et al, 2007; Duarte et al, 2007; Thiele et al, 2005), signaling networks (Papin and Palsson, 2004; Li et al, 2009), and more recently for macromolecular synthesis (Thiele et al, 2009, 2010). COBRA has been successfully applied for metabolism and enabled many biotechnological and biomedical issues to be address (Feist and Palsson, 2008; Oberhardt et al, 2009; Feist et al, 2009). So far, key questions of molecular evolution could not be investigate with these networks as they do not explicitly account for genes and proteins in a sequence-specific manner.

The most recent metabolic reconstruction of *E. coli* accounts for function of 1260 metabolic genes, which represents almost 30% of the open reading frames (ORF) in *E. coli*’s genome (Feist et al, 2007). We recently constructed the first genome-scale, stoichiometric network of the transcriptional and translational (tr/tr) machinery of *E. coli* (Thiele et al, 2009). The reconstruction accounts for 303 gene products, including ribosomal proteins, RNA polymerase, tRNA and rRNA. It represents the synthesis of all known components necessary to produce themselves. Here, we integrate these two reconstructions into a Metabolic-Expression (‘ME’) matrix reconstruction that accounts for the synthesis of almost 2,000 *E. coli* genes (SI Figure 1). To-date, only few examples of integrated networks of cellular functions have been published, including i) a metabolic-regulatory network using metabolic reconstruction and transcriptional regulatory network in form of Boolean expressions, for *E. coli* (Covert et al, 2004); and ii) metabolic-signaling-regulatory models (Covert et al, 2004; Lee et al, 2008). However, these integrated functional networks
do not explicitly account for proteins (enzymes and regulators) and they employ other modeling tools than COBRA (e.g., ordinary differential equations or Boolean logic). Therefore, the presented integrated network is unique and the first of its kind. Models derived from the ME-matrix reconstruction will allow us to address a new dimension of biotechnological, biomedical, and evolutionary questions that have not been modeled yet, including codon usage, protein engineering and prediction of cellular proteome. This ME-matrix formalism represents a milestone towards cell-scale modeling and sets stage in modeling techniques to achieve this ambitious goal in near future.

Results and discussion

We developed an integrated stoichiometric model of E. coli MG1655’s metabolic (Feist et al., 2007 and macromolecular synthesis machinery (Thiele et al., 2009) networks (Figure 1A). It accounts for the synthesis, assembly and function of 1,827 protein coding and 110 RNA coding genes corresponding to >40% of the genome at single nucleotide resolution. This model of metabolism and gene expression (ME-matrix) encompasses many cellular functions detailed in 76,589 reactions and 62,212 components (Figure 1D, Table 1). Transcriptional regulation, replication, signal transduction and transcriptional/ translational abortion events are outside the scope of this reconstruction. Many of the ME-matrix genes are highly conserved in enterobacter and non-enterobacter species. This conservation is particularly interesting as the ME-matrix accounts for all major antibiotic targets, except DNA gyrase (Kohanski et al., 2010), which could be exploited for functionally assessing lethal or sub-lethal antibiotic doses and combination therapies of novel antibiotic substances.

Metabolism provides precursors for the macromolecular synthesis machinery, which in turn synthesizes the metabolic enzymes that catalyse biochemical reactions (Figure 1A). An approximation of this interdependency results in linear inequalities that tightly constrained quasi-steady state reaction rates over many orders of magnitude (i.e., $\text{mmol}\cdot\text{g}^{-1}\cdot\text{hr}^{-1}$ vs. $\text{nmol}\cdot\text{g}^{-1}\cdot\text{hr}^{-1}$). The stoichiometric coefficients in the ME-matrix are distributed over four orders of magnitude, since many precursors are required to form one macromolecule (Figure 1B). We explicitly included enzymes into metabolic reactions; thus, additional constraints were needed to ensure that the catalyzing enzyme is synthesized within the network when its metabolic reaction is used (Material & Methods). We added linear coupling constraints that ensure that if an enzymes’ biosynthetic flux is zero then its utilization flux is zero (Figure 1C). Additionally, if an utilization reaction carries a high flux that the biosynthetic flux needs to also be higher (Thiele et al., 2010) (Material & Methods). The ME-matrix model is amenable to flux balance analysis (FBA) (Savinell and Palsson, 1992), the linear optimization of a biologically motivated objective (e.g., growth rate) subject to constraints (e.g., environmental conditions) (Price et al., 2004).

To assess the predictive potential of the ME-matrix, we compared the growth predictions with experimental data (SI Text). Metabolic models generally contain a biomass reaction, which stochiometrically weights the contribution of metabolic precursors towards synthesis of a new cell. The ME-matrix contains an adjusted biomass with reduced amino acid and nucleotide precursor requirement as they are used to produce the ME-matrix’ transcriptome and proteome (see SI Text for details). The ME-matrix reliably predicted quantitative growth rates in four well defined environmental conditions and outperformed the metabolic network (Figure 2 A). We also obtained
good agreement with experimental data reporting growth capability in 170 defined minimal media (Figure 2B and SI Text). These results gave us confidence in the ME-matrix’ growth phenotype prediction in a broad range of conditions.

We also wished to validate the ME-matrix’ predictive potential for the computed transcriptome. Therefore, we determined the effective in silico codon usage by calculating the codon usage for all ME-matrix genes and multiplying them with their average translation rate across 170 environmental conditions. The effective in vivo codon usage was determined similarly by multiplying the codon usage by the average expression level for each gene across various environmental and genetic conditions (Lewis et al., 2009). We found a high correlation between in silico and in vivo codon usage (Pearson correlation, \( R^2 = 0.901 \)) (Figure 2C). These results suggest that our predicted codon usage is physiologically relevant.

A key interest of systems biology is to develop a mechanistic basis for the genotype-phenotype relationship. The ME-matrix explicitly captures the nucleotide sequence for >1900 genes and stoichiometrically represents their cellular functions, so we asked if and how CUB (genotype) evolves to maximize growth rate (phenotype) in different growth environments. We generated a range of perturbed ME-matrices differing only in codon usage from the wildtype ME-matrix; ten ME-matrices with more biased codon usage ("biased strains", B1-B10) and five ME-matrices with less biased codon usage ("equilibrated strains", EQ1-EQ5) (Figure 3A, B). As expected, CUB of equilibrated strains was highly correlated while the CUB was idiosyncratic in biased strains (SI, Figure S6). With FBA, we calculated the strains’ growth rates across 170 conditions. The growth rates of the biased strains were comparable to the wildtype, while the equilibrated strains grew slower (Figure 3C). We identified five cases, where a biased strain was able to grow, but the wildtype was unable to grow (Figure 4B). Thus, these strains exhibited an increased fitness in these conditions. We also calculated growth rates using experimentally measured carbon and oxygen uptake rates as boundary conditions (Fong et al., 2005; Covert et al., 2004) (Figure 2A). Seven biased strains exhibited up to 50\% reduction in growth rates in two conditions, while the equilibrated strains showed reduced growth rates in all four conditions (Figure 3B). These results demonstrate that, in silico, changes in CUB can alter a strain’s ability to grow in certain environments and affect the growth rate. How is this possible given that perturbation to codon usage only affects stoichiometric coefficients and not the sparsity pattern of the perturbed ME-matrices as no reactions were added or removed?

Statistical genomic studies have identified GC content as the single most informative determinant of CUB (Chen et al., 2004; Knight et al., 2001). A recent study showed that CUB, but not GC content, correlated with minimum generation time (Vieira-Silva and Rocha, 2010). We examined the relationship between GC content, CUB and growth rate in wildtype and perturbed ME-matrices. Generally, we found that GC content correlated with growth rate, but there was a plateau at a wildtype growth rate (Figure 4A). At higher GC content, CUB seems to dominate the calculated maximal possible growth rate. The GC content of perturbed strains was between 45\% and 55\%, which is similar to the wildtype’s (53\%), therefore, the metabolic changes due to GC content changes were minor. An entropic measure of CUB indicates that the most biased strains have a significant lower growth rate than wildtype and equilibrated strains (SI, Figure S6).

The chief advantage of using the ME-matrix is the ability to derive causal hypotheses by analyzing numerical properties of flux balance analysis (FBA) solutions. With FBA, the amount by which the growth rate would increase with 1 \( \text{nmol.g}^{-1}\text{DW.hr}^{-1} \) relaxation of bounds on each reaction rate, is given by the optimal reduced cost (RC) returned by a linear programming algorithm. Each
mutant ME-matrix had bounds identical to the wildtype: i) upper bounds on each transcription (initiation) reaction, i.e., RNA polymerase elongation rate times gene dosage, ii) upper bounds on the sum of tRNA synthesis rates, iii) an upper bound on uptake rates of carbon source and/or oxygen, and iv) a lower bound on non-growth associated ATP maintenance. The wildtype was growth limited by upper bounds on ribosomal RNA (rRNA) operon transcription reactions, as indicated by the largest RCs. The same bounds limited growth of biased strains in glycerol and glucose (anaerobic) growth conditions. In contrast, in lactate and glucose (aerobic) conditions, the bounds on leucyl-tRNA transcription reactions were limiting growth in all seven growth impaired biased strains (Figure 4C). Leucine is the most abundant amino acid in E. coli’s genome, which encodes for eight leucyl-tRNAs, five of which can read the most frequent codon CTT. In all biased strains with reduced growth rate, the change in CUB shifted CTT to minor ones, which are read by single tRNA species (Figure 4C, inset). Five biased strains had a highest RC for the \( tRNA_{\text{leuU}} \) transcription reaction. In the case of strain B2, the growth limiting tRNA species was condition dependent. We numerically confirmed that relaxation of bounds corresponding to the highest RC was sufficient to restore wildtype growth rate and rRNA transcription as the growth limiting bound. The growth rates of equilibrated strains were limited to a similar degree by bounds on many different tRNA transcription reactions. tRNA supply was growth limiting in the perturbed ME-matrices.

In the wildtype, rRNA, but not tRNA, transcription was growth limiting in the tested environments. This result is in agreement with experimental data reporting correlation between ribosome number and growth rate (Bremer and Dennis, 1996). The wildtype CUB and tRNA supply must therefore be complimentary to the tRNA demand for each tested environment. Reduced or no growth of a perturbed ME-matrix in an environment was caused by an imbalance of demand-and-supply of tRNA species, as not all proteins to sustain growth could be synthesized (Figure 4C). The identification of growth enabling, perturbed ME-matrices (Figure 4B) suggests that the wildtype operon structure was consistent with CUB of co-expressed genes in most but not all environments. Our results indicate that CUB reflects environments an organism can occupy, which agrees with statistical genomic studies (Willenbrock et al, 2006; Vieira-Silva and Rocha, 2010).

Upon CUB perturbation, an increased tRNA demand may be met by augmenting supply that could be achieved by i) genome re-organization to relocate tRNA genes closer to the origin of replication, which would increase the gene copy number via gene dosage effect; ii) acquisition of tRNA genes from other organisms; or iii) modification of a tRNA to expand its set of read codons. It has been recently that a second leucyl-tRNA (\( tRNA_{\text{leuW}} \)) is able to read CTT in E. coli MAS39 due to a uridine-5-oxyacetic acid modification (Sorensen et al, 2005), which we have not accounted for in the current ME-matrix. It remains to be established that \( tRNA_{\text{leuW}} \) in E. coli MG1655 can also read CTT. Pinpointing a lack of robustness in the perturbed ME-matrix highlights its use to derive new hypothesis.

**Conclusion**

Network reconstruction technologies developed over the past 10 years enabled us to build an integrated metabolic, macromolecular synthesis reconstruction for E. coli K12 MG1655. This ME-matrix is a knowledge-base and it can also be used for computations enabling the simultaneous reconciliation of the activities of its gene products. We introduced perturbations to CUB that affected all ME-matrix genes equally thereby creating mutant strains that satisfy the mutational hypothesis.
Observed changes in growth rate of most strains were caused by an unattainable proteome to meet metabolic requirements due to tRNA supply shortage. The adjustment of this shortage through expansion of tRNA content or reading is most consistent with the natural selection hypothesis. Our results strengthen also previous observations that synonymous codon usage significantly impacts achievable growth phenotypes (Bragg and Wagner, 2009) as we identified reduced and increased maximal growth rates of the in silico strains depending on environmental niche. Using a genome-scale analysis framework that is novel to molecular systems biology, we provided an explanation on how expansion of tRNA content and/or reading could be used as an evolutionary mechanism to deal with mismatches between CUB (genotype) and environment to maximize growth rate (phenotype).

Materials and methods

Metabolic reconstruction. The metabolic reconstruction of E. coli, iAF1260 (Feist et al, 2007), was obtained in SBML format (Ec_iAF1260_flux1.xml), from http://systemsbiology.ucsd.edu/In_Silico_Organisms/E_coli/E_coli_SBML) and imported into Matlab (Mathworks Inc.) using the COBRA Toolbox (Becker et al, 2007). iAF1260 accounts for 1,260 E. coli genes and 2,077 reactions, including 1,339 unique metabolic reactions, 690 transport reactions, and 304 exchange reactions (Feist et al, 2007). 1,294 reactions have gene-protein-reaction associations. iAF1260 accounts for 1,039 unique metabolites. A total of 1,148 unique, functional proteins are accounted for including 167 multigene complexes and 346 isozymes (Feist et al, 2007). Prior to merging with the 'E-matrix', all gene associations connected to the artificial gene 's0001' were removed. iAF1260 contains tRNA charging reactions that were also removed from the model before integration.

Macromolecular synthesis reaction. The tr/tr machinery reconstruction, deemed expression or 'E-matrix', was imported into Matlab (E_matrix.mat) (Thiele et al, 2009). It accounts for 249 transcription units containing 423 genes, 228 proteins (34 without coding gene), 86 tRNA species, 22 rRNA species, and one miscellaneous RNA species. A total of 11,991 network components and 13,694 reactions describe the synthesis, assembly, and function of the macromolecular synthesis machinery of E. coli K12 MG1655. The bounds on exchange and transport reactions for metabolites, which were present in the E-matrix and in iAF1260, were set to be zero (lower and upper bound) in the E-matrix prior to integration.

A detailed explanation of ME matrix reconstruction can be found in the SI Text. The COBRA approach has been described in detail elsewhere (Palsson, 2006; Orth et al, 2010) and a short description can be found in the SI Text.

Clusters of orthologous groups (COGs). The functional gene coverage included in the ME-reconstruction may be best accessed by looking at the distribution of COGs (Tatusov et al, 2003). A total of 2,806 E. coli’s genes have an assigned COG function, of which 1,436 are in the ME-reconstruction. The remaining 496 ME-genes have no COG information and thus cannot be considered for functional coverage analysis. The transcription category contains 142 genes that are currently not included in the ME-matrix, as it does not account for transcriptional regulation yet. Similarly, genes of the replication and signal transduction categories are missing.

Adjustment of biomass. The amino acid and growth associated maintenance (GAM) of the E. coli biomass reaction in the ME-matrix was adjusted to account for the cost of synthesis of the machinery and proteins in the ME-matrix. After performing a sensitivity analysis for these
two parameters (see Result section), we adjusted the biomass reaction to account for 50% of the amino acid content and 50% of the GAM of the biomass reaction in the metabolic reconstruction, iAF1260. The adjusted biomass reaction was used in all simulations if not noted differently.

**Coupling constraints.** The conversion of a reconstruction to a mathematical model normally consists of the definition of the systems boundaries, the addition of exchange and demand reactions and the application of condition-specific constraints on exchange and/or intracellular reactions (Thiele and Palsson, 2010). The same steps were undertaken to convert the ME-reconstruction into ME-matrix models. However, an additional step needed to be performed, which added further constraints to the model. These constraints are called coupling constraints and link (or “couple”) the flux through a biosynthetic flux, $v_s$, (e.g., transcription) with the corresponding utilization reaction(s), $v_4$, (e.g., translation) (Figure 1C). The formulation of the constraints ensures that if a biosynthetic flux is zero then its utilization flux has to be zero as well. An upper bound on the coupling constraint ensured that if a utilization reaction carries a high flux that the biosynthetic flux is higher as well. This requires the network to produce more gene products if they are highly used and thus represents a limit on enzyme capacity. The coupling constraints and their consequences to the steady-state flux solution space have been discussed in detail in (Thiele et al, 2010). Coupling constraints were included between: i) transcription and translation, ii) translation and protein utilization (e.g., flux through metabolic reaction), and iii) tRNA synthesis and tRNA utilization. These linear inequality coupling constraints retain the numerically scalable character of flux balance analysis. See Thiele et al. (Thiele et al, 2010) and SI Text for more details.

**Simulation constraints.** Experimental measurements of substrate and oxygen uptake rates were applied on the exchange reactions (Figure 2A). Note that the unit of the ME-matrix is nmol$ \cdot g^{-1} \cdot hr^{-1}$. Therefore, the listed rates were multiplied by a factor of $10^6$. In addition, the maximal reaction rates of stable RNA synthesis were constrained as described in (Thiele et al, 2009). Similarly, the maximal reaction rates of mRNA synthesis were constrained using the same approach but changing the mRNA transcription elongation rate (Bremer and Dennis, 1996). In all simulations, the non-growth associated maintenance (ATPM) requirement was set to $v_{\text{min,ATPM}} = v_{\text{max,ATPM}} = 8.39 \cdot 10^6$ nmol$ \cdot g^{-1} \cdot hr^{-1}$ as defined in (Feist et al, 2007).

**Growth comparison with Biolog and iAF1260.** Biolog data were downloaded from the website (http://biolog.com) for E. coli K12 MG1655. A total of 170 tested compounds were in the reconstruction. The different environments were simulated by adding compounds to a base medium and allowing oxygen to be consumed ($v_{\text{min}} = -18.5 \cdot 10^6$ nmol$ \cdot g^{-1} \cdot hr^{-1}$ and $v_{\text{max}} = 0$ nmol$ \cdot g^{-1} \cdot hr^{-1}$).

The base medium allowed the free uptake of specific compounds by setting their corresponding lower bound to $v_{\text{min}} = -1 \cdot 10^9$ nmol$ \cdot g^{-1} \cdot hr^{-1}$ (SI Text).

Furthermore, the maximal possible transcription rates for each stable RNA transcription unit and for each protein coding gene were limited assuming a doubling time of 24 minutes (which provides an upper bound), since we have no information about growth rates for the different growth conditions tested in the Biolog setup.

The ribosome production rate (DM_rib_50) and the biomass reaction (Ec_biomass_iAF1260_core_59p81M) were unbounded. Each nutrient was added to the base medium by setting the corresponding uptake rate to $v_{\text{min}} = -10 \cdot 10^6$ nmol$ \cdot g^{-1} \cdot hr^{-1}$ in the case of carbon sources, and $v_{\text{min}} = -20 \cdot 10^6$ nmol$ \cdot g^{-1} \cdot hr^{-1}$ in the case of nitrogen, phosphorus, and sulphur sources. Default elemental sources were as follows: D-glucose as carbon source, ammonium ion (NH$_4^+$) as nitrogen
source, orthophosphate ($p_i$) as phosphorus source, and $SO_4$ as sulfur source. The sources were added to the base medium, when the corresponding source was not tested for. The growth results for iAF1260 were obtained from (Feist et al, 2007).

Creation of in silico strain library. The biased strains and the equilibrated strains were produced as described in the SI Text. Briefly, the CUB was perturbed as illustrated in Figure 3A by replacing a codon by one of the possible synonymous codons either i) randomly resulting in biased strains, or ii) such that every codon has equal usage resulting in equilibrated strains. The replacement was repeated 100 times. The resulting gene sequences were stored. Each perturbed ME-matrix had its own stoichiometric matrix. The change of codon usage was introduced to ME-matrix by (i) adapting the nucleotide triphosphate requirements in the corresponding transcription reactions, (ii) changing the nucleotide monophosphates released in the mRNA degradation reactions, and (iii) updating the tRNA species according to the new codons (SI, Table S3, S4). Neither start codon nor stop codons were modified in the strains. For each perturbed ME-matrix and the wildtype ME-matrix, we applied the same simulation constraints and compared the maximal computed growth rate.

Numerical tests. Calculating with the ME-matrix is rather time-consuming and numerically challenging due to the matrix’s stiffness (Figure 1B). Therefore, it is required to test each computed point if it lies within the solution space (i.e., test if $S \cdot v = 0$ is true). We evaluated every solution for feasibility status returned by the solver and the associated error ($S \cdot v = 0$).

All simulations were carried out in Matlab (Mathwork, Inc.) using Tomlab (Tomlab, Inc.) as numerical analysis interface for linear programming.

The ME-matrices used in this study will be available under http://notendur.hi.is/ithiele/downloads.html.

Acknowledgements

This work was supported by a grant from the Department of Energy (DE-SC00092009), by a grant from the National Institutes of Health (R0157089), and by a Marie Curie International Reintegration Grant awarded to IT (249261) within the 7th European Community Framework Program. The authors declare that they have no competing financial interests.

Author Contributions

IT and BOP designed research. IT performed research. IT, RQ, AB, and RMTF contributed analytic tools. IT and RMTF analyzed data. IT, RMTF, and BOP wrote the paper.
Figure 1: **Overview of the ME-matrix.** A: Functional synergy between metabolism and macromolecular synthesis. B: ME-matrix histogram of stoichiometric coefficients (left) and sparsity pattern (right). The stoichiometric coefficients are spread over five orders of magnitude because of the difference in biochemical moieties required for metabolic and macromolecular synthesis reactions. C: Coupling constraints were be added to the ME-matrix to link (or "couple") the flux through a biosynthetic flux, $v_s$, (e.g., transcription) with the corresponding utilization reaction(s), $v_4$, (e.g., translation) (Thiele et al, 2010). $v_D = \text{protein dilution}$. D: Distribution of clusters of orthologous groups is shown for a total of 2,806 *E. coli* genes, of which 1,436 are in the ME-matrix.
Figure 2: **Figure 2: ME-matrix validation.**  
**A:** Comparison of predicted and experimentally determined growth rates (Covert et al., 2004; Fong et al., 2005). SUR = substrate uptake rate. OUR = oxygen uptake rate. $O^+$ = aerobic. $O^-$ = anaerobic. WT = wildtype. EV = evolved strain.  
**B:** Comparison of qualitative growth phenotype data with predicted *in silico* growth phenotype of ME-matrix (ME) and of metabolic model (iAF) prediction across 170 environments.  
**C:** Correlation between effective *in silico* and *in vivo* codon usage.
Figure 3: **Properties of in silico strains.**  
A: Differences between CUB of wildtype and perturbed ME-matrices.  
B: Heatmap of the usage of 61 codons (including start codon) in wildtype and mutant strains.  
C, D: Relative growth rates achieved by in silico strains across 170 environmental conditions (C) and when measured SUR and OUR were chosen as constraints (D).
Figure 4: **Distinguishing features between the in silico strains.**

A: GC content versus *in silico* growth rates. B: Growth rates in conditions, where the wildtype could not grow but biased strains and the metabolic model, iAF1260 (Feist *et al.*, 2007), did. C: RC results are shown for the biased strains in the four defined growth conditions. Increasing the flux rate through the transcription reaction of the tRNA transcription units (e.g., tscr_in TU00518_stab encoding $tRNA^{LeuU}$) by 1 $nmol.g_{DW}.hr^{-1}$ would increase the growth rate by 0.307 $mmol.g_{DW}.hr^{-1}$ (B1). The RCs are given in $mmol.g_{DW}.hr^{-1}$. Inset: changes of major leucine codon in the CUB perturbed ME-matrix. tRNA recognizing the codon is given in parenthesis. Leu1 and Leu2 are generic tRNA species representing multiple leucyl-tRNA species, see SI, Table S3,S4 and (Thiele *et al.*, 2009).
### Tables

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription Units</td>
<td>1,152</td>
</tr>
<tr>
<td>Genes</td>
<td>1,937</td>
</tr>
<tr>
<td>- Protein coding genes</td>
<td>1,827</td>
</tr>
<tr>
<td>- RNA coding genes</td>
<td>110</td>
</tr>
<tr>
<td>Network reactions</td>
<td>76,589</td>
</tr>
<tr>
<td>Network components</td>
<td>62,212</td>
</tr>
<tr>
<td>Coupling constraints between reactions</td>
<td>3,044</td>
</tr>
</tbody>
</table>

Table 1: **ME-matrix statistics.** Summary of ME-matrix content.
References


Fong SS, Joyce AR, Palsson BØ (2005) Parallel adaptive evolution cultures of *Escherichia coli* lead to convergent growth phenotypes with different gene expression states. *Genome Research* **15**: 1365–1372


Supplementary Information

Materials and Methods

Reconstruction of the 'ME'-matrix

Constraint-based reconstruction and modeling approach

The reconstructed biochemical network is often represented in a tabular format, listing all network reactions and metabolites in a human-readable manner (see (Thiele and Palsson, 2010) for details). The conversion into a mathematical, or computer-readable format, can be done automatically by parsing the stoichiometric coefficients from the network reaction list (e.g., using the COBRA toolbox (Becker et al., 2007)). The mathematical format is called a stoichiometric matrix, or $S$ matrix, in which the rows correspond to the network metabolites and the columns represent the network reactions. For each reaction, the stoichiometric coefficients of the substrates are listed with a minus sign in the corresponding cell of the matrix, while the product coefficients are positive numbers. The resulting size of the $S$ matrix is $m \times n$, where $m$ is the number of metabolites and $n$ the number of network reactions. Mathematically, the $S$ matrix is a linear transformation of the flux vector

$$v = (v_1, v_2, \ldots, v_n)$$

to a vector of time derivatives of the concentration vector

$$x = (x_1, x_2, \ldots, x_m)$$

as

$$\frac{dx}{dt} = S \cdot v$$

At steady-state, the change in concentration as a function of time is zero; hence, it follows:

$$\frac{dx}{dt} = S \cdot v = 0$$

The set of possible flux vectors $v$ that satisfy this equality constraint might be subject to further constraints by defining

$$v_{i,\text{min}} \leq v_i \leq v_{i,\text{max}}$$

for reactions $i \in n$. In fact, for every irreversible network reaction $i$, the lower bound was defined as $v_{i,\text{min}} \geq 0$ and the upper bound was defined as $v_{i,\text{max}} \geq 0$. Exchange reactions supply the network with nutrients or remove secretion products from the medium. The uptake of a substrate by the network was defined by a flux rate $v_i < 0$ and secretion of a by-product was defined to be $v_i > 0$ for every exchange reaction $i$. Finally, the application of constraints corresponding to different environmental conditions (e.g., minimal growth medium) or different genetic background (e.g., enzyme-deficient mutant) allow the transition from biochemical network reconstruction to condition-specific model. Note that the network reconstruction is unique to a target organism (and defined by its genome) while it can give rise to many different models by applying condition-specific constraints. In this study, all flux rates are given in $\frac{\text{nmol}}{\text{g DW} \cdot \text{h}}$ if not stated differently.

Construction of transcription and translation reactions for metabolic enzymes

The integration of the E-matrix with the iAF1260 requires that all metabolic enzymes (1260 gene products) are synthesized by the network. Therefore, we used the template reactions for transcription, translation, mRNA degradation, etc. as well as the gene information (e.g., transcription unit assignment from EcoCyc (Karp et al., 2004), gene coordinates and gene direction from (Riley et al., 2006)) (see Table 2 for a complete list). The formulation of the reactions was done in an automated fashion as described elsewhere (Thiele et al., 2009).
Protein Complex formation

Information about protein complex formations was obtained from iAF1260, which describes the relationship between gene products and metabolic reactions in terms of Boolean logic (Feist et al., 2007). This information was complemented with protein complex formation information obtained from EcoCyc (Karp et al., 2004) and primary literature. Protein complex formation reactions for multimeric proteins were formulated manually assuming that all subunits bind simultaneously. A monomeric subunit was assumed when no information was available.

Metallo-ions and prosthetic groups

Information about metallo-ion and/or prosthetic groups were obtained from EcoCyc (Karp et al., 2004), protein structures of E. coli enzymes and primary literature. The information was manually assembled, while the network reactions were formulated based on the template reactions (see Thiele et al., 2009 for details). If no information about the number of associated ions could be found, we assumed one ion per monomer.

Creation of ME-matrix

The tr/tr reactions for all metabolic genes were added to the E-matrix by adding additional rows and columns for the new components and reactions, respectively. The integration of iAF1260 and this extended E-matrix was done computationally by creating a non-redundant reaction list containing the union of both reconstructions. Functional overlap between the two networks exists on two points: i) exchange reactions of the E-matrix and the metabolic synthesis reactions; and ii) the metabolites incorporated by the E-matrix into RNA and proteins that are also consumed by the biomass reaction of the metabolic network (Figure Main Paper 1A). The metabolic reactions were reformulated to include their enzymes in a subsequent step, which resulted in the ME-matrix. The ME-matrix generation involved adding enzymes, enzyme complexes, and inactive enzymes to the reactions.

Consider the following reaction (G6PP):

\[ \text{G6P} + H_2O \rightleftharpoons \text{Glc} - D + P_i \]

The preceding equation can be changed by adding enzymatic complexes. First, information is collected about the reaction (G6PP):

- Gene loci = b0822
- Gene = ybiV
- Protein = YbiV

Second, the original reaction is converted to the following (notice name change):

\[ \text{G6PP}_A : \text{G6P} + H_2O + YbiV_{\text{mono}} \rightleftharpoons YbiV_{\text{G6P\_cplx}} \]

Third, new reactions are added to the end of the reaction list.

\[ \text{G6PP}_B : YbiV_{\text{G6P\_cplx}} \rightarrow YbiV_{\text{Glc} - D\_cplx} \]
\[ \text{G6PP}_C : YbiV_{\text{Glc} - D\_cplx} \rightarrow \text{Glc} - D + P_i + YbiV_{\text{mono\_inact}} \]
\[ \text{G6PP}_DREC : YbiV_{\text{mono\_inact}} \rightarrow YbiV_{\text{mono}} \]

If a reaction is reversible (which G6PP is), the reverse reactions are:

\[ \text{G6PP}_E : \text{Glc} - D + P_i + YbiV_{\text{mono}} \rightleftharpoons YbiV_{\text{Glc} - D\_cplx\_R} \]
\[ \text{G6PP}_F : YbiV_{\text{Glc} - D\_cplx\_R} \rightarrow YbiV_{\text{G6P\_cplx\_R}} \]
\[ \text{G6PP}_G : YbiV_{\text{G6P\_cplx\_R}} \rightarrow \text{G6P} + H_2O + YbiV_{\text{mono\_inact}} \]
If the equation occurred in the periplasm ([p]) or extracellular space ([e]), transport reaction was also included. The reaction G6PP is located in the cytoplasm, not requiring transport reactions, however if hypothetical transport reaction(s) would be as follows:

- If in the periplasm:
  \[ YbiV_{\text{export}}[p] : YbiV_{\text{mono}} \rightleftharpoons YbiV_{\text{mono}}[p] \]
- If in the extracellular space:
  \[ YbiV_{\text{export}}[p] : YbiV_{\text{mono}} \rightleftharpoons YbiV_{\text{mono}}[p] \]
  \[ YbiV_{\text{export}}[e] : YbiV_{\text{mono}}[p] \rightleftharpoons YbiV_{\text{mono}}[e] \]

When the reaction is in either the periplasm or extracellular space, the protein \( YbiV_{\text{mono}} \) would change to be \( YbiV_{\text{mono}}[p] \) and \( YbiV_{\text{mono}}[e] \), respectively. Note that we do not capture any protein secretion mechanisms in the ME-matrix.

All lower and upper bounds are set to the -inf and +inf, unless the reaction is only in the forward direction (0 +inf).

The aforementioned example assumes only one gene to one protein relationship. There are three other possibilities.

First is the "OR" case. Two or more different genes can code for a protein, which can catalyze the same reaction. In this case, each gene is treated as its own reaction as shown above. Therefore, if the G6PP reaction could be catalyzed by \( YbiV \) and some other protein (\( XxyY \)), the script would create the reactions listed above and also repeat the process with \( XxyY \) reactions. In this instance, the naming convention for reactions is changed. Instead of using \( \text{G6PP}_A \), \( \text{G6PP}_{YbiV}_A \) and \( \text{G6PP}_{XxyY}_A \) are used to differentiate between the different catalyzing enzymes.

The second instance is the "AND" case. Multiple genes code multiple proteins that must form a complex to catalyze the reaction. In this case, an additional reaction is added known as the complex formation reaction. Suppose \( YbiV \) and \( XxyY \) are both required for \( \text{G6PP} \). A complex formation reaction would be created:

- \( YbiV_{-XxyY_{\text{cplx}_{\text{FORM}}}} : YbiV + XxyY \rightleftharpoons YbiV_{-XxyY_{\text{cplx}}} \)

This new complex would then be used in the reactions above replacing \( YbiV_{\text{mono}} \).

The third instance is the combination of both the "OR" and "AND" case. The rules laid out above are then used to combine the two.

Composition of base medium

The base medium allowed the free uptake of the following compounds by setting their corresponding lower bound to \( v_{\text{min}} = -1 \times 10^9 \frac{\text{nmol}}{\text{gDW} \cdot \text{h}} \):

- \( \text{EX}_h2\text{s(e)} \)
- \( \text{EX}_\text{ca2(e)} \)
- \( \text{EX}_\text{cl(e)} \)
- \( \text{EX}_\text{co2(e)} \)
- \( \text{EX}_\text{cobalt2(e)} \)
- \( \text{EX}_\text{cu2(e)} \)
- \( \text{EX}_\text{fe2(e)} \)
- \( \text{EX}_\text{fe3(e)} \)
- \( \text{EX}_\text{h2o(e)} \)
- \( \text{EX}_h(e) \)
- \( \text{EX}_k(e) \)
- \( \text{EX}_\text{mg2(e)} \)
- \( \text{EX}_\text{mn2(e)} \)
- \( \text{EX}_\text{mobd(e)} \)
- \( \text{EX}_\text{na1(e)} \)
- \( \text{EX}_\text{tungs(e)} \)
- \( \text{EX}_\text{zn2(e)} \)
- \( \text{EX}_\text{cbl1(e)} \)

Coupling constraints

There are three dominant sets of constraints applied to the ME matrix.

1. Constraints on the exchange reactions to simulate different environmental conditions (see below).

2. Constraints on the maximal transcription rate for stable and messenger RNA
   - these two set of constraints are on \( v_i \): \( v_{i,\text{min}} \leq v_i \leq v_{i,\text{max}} \)
3. Coupling constraints on reactions, in the form of

\[ v_4 - c_{\text{min}} \cdot v_S \geq -s, \quad s \geq 0 \]  
(1)

\[ v_4 - c_{\text{max}} \cdot v_S \leq 0 \]  
(2)

where \( c_{\text{min}} \) and \( c_{\text{max}} \) are the bounds on the proportion factor or 'coupling coefficient'. \( s \) can be used to allow accumulation of a network component without using it in the steady-state solution thereby relaxing the requirement that all synthesized network components need to be used within the network. Eq. 2 ensures that a higher flux through \( v_4 \) raises the lower bound on the synthesis reaction \( v_S \). These linear inequality coupling constraints retain the numerically scalable character of flux balance analysis.

Coupling constraints were added as described in the main text and in (Thiele et al., 2010):

- between mRNA synthesis and translation (via _mRNA_degr1 and _CONV2):
  - based on equation 1 and 2 the coupling is achieved. The parameters \( c_{\text{min}} \) and \( c_{\text{max}} \) are determined based on the following equation
    \[
    \frac{r_{\text{tl}}}{L_{\text{prot},i}} \cdot T_{\frac{1}{2},i} \cdot 3600 \log(2)
    \]  
(3)

    , where \( r_{\text{tl}} \) is the translation rate at a given doubling time \( (T_d) \), \( L_{\text{prot},i} \) is the length of the protein \( i \) (in amino acids), and \( T_{\frac{1}{2},i} \) is the half-life time of mRNA \( i \). The upper bound on half-life time was assumed to be 60 minutes, while the lower bound was set to be 0.1 minute.

- between protein synthesis and protein utilizing reactions
  - for E matrix proteins: via DM_ and _RECYCL
    
    * based on equation 1 and 2 the coupling is achieved. The parameters \( c_{\text{min}} \) and \( c_{\text{max}} \) are were set to 1 and 10,000, respectively. This parameter corresponds to the turnover/utilization rate of the protein. Note, that the proteins are currently not degraded in the network since they are considered to be stable (at least in respect to the time span of the doubling times considered here).

  - for M matrix proteins (metabolic enzymes): via DM_ and _DREC
    
    * The rational behind the parameters \( c_{\text{min}} \) and \( c_{\text{max}} \) is quite similar to the E-matrix proteins.
    * In general we observed that the \( c_{\text{max}} \) should be chosen higher for enzyme as the metabolic rates are much higher than the synthesis reaction rates and thus either more protein are needed or a higher activity (utilization rate) is needed to meet this greater demand.

- between tRNA charging and utilization (see (Thiele et al., 2010))
Note that the demand reactions represent the “real” accumulation of proteins observed in cells (and measured by proteomic approaches). Hence, the proteomic data can be directly mapped onto the demand reactions. The situation is quite different for the transcripts as a real turnover rate exists and no “real” accumulation is possible. A temporary accumulation is modeled via _CONV2 reactions, which allow each mRNA species to cycle within the network for a determined time before getting turned over. Hence, the mRNA synthesis corresponds to the maintenance of the internal pool for each transcript within the cell. In fact equation 3 controls the pool size for each transcript within the network. There is a direct conversion possible from _CONV2 and _mRNA_degr1 allowing to back-calculate half-life time and mRNA concentration for each transcript given a flux vector (Thiele et al., 2010).

**Single gene deletion study of metabolic enzymes**

The ME-matrix accounts for 1,260 metabolic genes. We tested the *in silico* growth phenotype of the single gene deficient strain in aerobic glycerol minimal medium and compared them with recently published experimental study (Joyce and Palsson, 2006) and with the *in silico* single knockouts of iAF1260 (results were taken from (Feist et al., 2007)).

Performing a single gene deletion study in the ME-matrix is a little different to the single deletion study in metabolic networks, because (i) proteins are explicit part of the metabolic reactions and (ii) transcription may occur with other genes (if co-expressed in a transcription unit) and thus coupling constraints would cause all genes in the transcription unit to not be expressed. Therefore, all translation initiation reactions for the gene are identified (e.g., 'tl_ini_bxxx’) and the corresponding lower and upper bounds were set to zero. Then, all coupling constraints were identified and removed (see above for more details on coupling constraints). We then maximized for the biomass reaction in the *in silico* knockout strain. The same procedure was repeated for all 1,260 metabolic genes.

**Creation of *in silico* strain library**

**Strains with biased codon usage**

The biased strains were generated using the following algorithm:

**Input:** model, sequence for each gene in model, number of iterations m

**Output:** model_biased

**Algorithm:**

1. Choose randomly a codon, $c_1$
2. Identify possible synonymous codons: $c_s = \{c_1 = c_{s1}, c_{s2}, ..., c_{sk}\}$
3. Choose randomly one codon from $c_s$: $c_{si}$
4. Replace all instances of $c_1$ with $c_{si}$
5. Update ME-matrix for all genes based on new gene sequence:
   (a) Transcription reactions
   (b) mRNA degradation reactions
   (c) Translation reactions (tRNA molecule will be updated based on codon recognition)
6. Repeat 1 through 5 m times
Strains with equilibrated codon usage
The equilibrated strains were produced as follows:

**Input:** model, sequence for each gene in model, number of iterations m

**Output:** model_eq

**Algorithm:**

1. Initialize vector codon = zeros, which will count the occurrences of different codons in the genome
2. Define a random order of genes to start step 3
3. For each gene \( i \) of the model genes
   (a) For each codon \( c_{s,j} \) in gene sequence \( i \)
   (b) Identify possible synonymous codons: \( c_{s,j} = \{ c_1 = c_{s1}, c_{s2}, ..., c_{sk} \} \)
   (c) Choose codon \( c_{s,j} \) from \( c_{s,j} \) with lowest usage in vector codon
   (d) Replace \( c_{s,j} \) with \( c_{s,j} \) in gene sequence \( i \)
   (e) Update codon
4. Update ME-matrix for all genes based on new gene sequence:
   (a) Transcription reactions
   (b) mRNA degradation reactions
   (c) Translation reactions (tRNA molecule will be updated based on codon recognition)
5. Repeat 1 and through 4 \( m \) times

In all case, \( m \) was 100. Note that each strain has its own ME-matrix, which contains the alterations in the \( S \) matrix but has the same reaction and coupling bounds as the wildtype.

The change of codon usage was introduced to ME-matrix by (i) adapting the nucleotide triphosphate requirements in the corresponding transcription reactions, (ii) changing the nucleotide monophosphates released in the mRNA degradation reactions, and (iii) updating the tRNA species according to the new codons (SI, Table S4). Note that neither start codon nor stop codons were modified in the strains.

**Supplementary results**

**Model Content**

**Reconstruction content**

The integrated reconstruction of metabolism and gene expression, ME-matrix, accounts for 1,260 metabolic genes, 303 macromolecular synthesis machinery genes, and 375 co-expressed genes without function in the ME-matrix. A total of 1,823 protein coding genes and 115 RNA coding genes are captured in the ME-matrix along with their synthesis reactions resulting in active, functional gene products. The generation of the synthesis reactions leading to functional gene products has been described in great detail elsewhere(Thiele et al., 2009).

**Properties of the components**

The ME-reconstruction accounts for information regarding protein complex formation, metallo-ion requirement, and necessary prosthetic groups. The metabolic reconstruction provided gene-protein-reaction (GPR) associations encoding via Boolean rules which gene products catalyze a metabolic function (Thiele and Pals son, 2010). While GPRs capture heteromeric complexes, they do not contain any information regarding homomers. A total of 495 protein complex formation reactions were added manually based on literature (Table 2). Furthermore, 305 proteins have covalently bound metallo-ions (Table 1). The ME-reconstruction also accounts for eleven different kinds of prosthetic groups in 99 proteins (Table 1). This latter information has not been considered in any other biochemical reconstruction.
GC content

The GC content of the individual strains was calculated by counting the instances of guanine and cytosine residues in the 1,823 protein coding genes included in the ME-matrix. The genome sequence used for this analysis was version m56, (Blattner et al., 1997).

Entropy

In order to quantify the degree of synonymous codon bias in a sequence we computed the synonymous codon entropy (Zeeberg, 2002). We used the entropy function since it reaches a maximum when all codons have equal probability of coding for their respective amino acids. Conversely, the entropy reaches its minimum when each amino acid is exclusively coded for by one of its possible codons. The synonymous codon entropy, \( H_{\text{synnon}} \), is defined as:

\[
H_{\text{synnon}} = - \sum_{a=1}^{20} \left( N_a \left( \sum_{c=1}^{64} p_{ac} \ln p_{ac} \right) \right) / \sum_{a=1}^{20} N_a
\]

where \( p_{ac} \) is the probability that amino acid \( a \) is encoded by codon \( c \), and \( \ln \) denotes the natural logarithm. If no amino acid is not coded for by a particular codon, \( p_{ac} = 0 \), then we use the definition \( 0 = 0 \ln 0 \). Here we weight the contribution to the total synonymous codon entropy by the number of each particular amino acid, \( N_a \), within a sequence. This means that a rare amino acid with highly biased synonymous codon usage does not overly effect the total entropy of a sequence if the remainder of the common amino acids have relatively unbiased codon usage. Since we wish to compare the synonymous codon bias between genes, we normalize the total by the total number of amino acids in a sequence, \( \sum_{a=1}^{20} N_a \). If we wish to calculate the total entropy for a set of genes then we simply sum up the synonymous codon entropy for each gene’s sequence, then divide by the total number of genes. Therefore, the total synonymous codon entropy is comparable between different sequences, such as mutant biased, wild type, and mutant equilibrated strains, which have low, medium and high total synonymous entropy, respectively, (Figure 6).

Model Validation

Adjustment of biomass

The ME-matrix accounts for the synthesis of almost half of the functions encoded in \( E.\ coli \)'s genome. Subsequently, the biomass reaction, which accounts for precursors to the macromolecular building blocks, needs to be adjusted for the fraction of amino acids (AA) and nucleotide triphosphates (NTP) used for synthesis of ME-matrix gene products. Therefore, we carried out a sensitivity analysis to identify the best parameters such that the model achieved the experimentally observed growth rates (Figure 2). Two main parameters were considered, the fraction of (i) amino acids and (ii) growth associated maintenance (GAM). The latter one is included in the biomass reaction to account for the energy necessary to synthesize RNA and proteins (in terms of ATP hydrolysis)(Thiele and Palsson, 2010; Feist et al., 2007). Note that we did not alter the fraction of NTPs since their overall contribution is relative small in the biomass reaction. We found that a good overlap between \( \textit{in silico} \) and \( \textit{in vivo} \) growth rate was achieved when the biomass reaction was adjusted to 50% of the amino acid requirement and 50% of the GAM. Finetuning of these two parameters may lead to an improvement of quantitative growth rate predictions.

Quantitative prediction of growth phenotypes

The ME-matrix can predict quantitative growth phenotypes given experimentally measured substrate and oxygen uptake rates (Figure Main Paper 2A). The experimental data were obtained from the literature and correspond to wildtype strains in multiple environmental conditions (i.e., minimal medium supplemented with glucose, glycerol, or lactate in aerobic and anoxic conditions). Furthermore, the wildtype cells were evolved on minimal medium supplemented with glycerol or lactate and after 60 days of evolution (with optimal growth as selection pressure) the substrate and oxygen uptake rates were measured (Fong et al., 2005; Covert et al., 2004). We compared the ME-matrix predictions with optimal growth rates calculated with iAF1260. We found that in many cases the metabolic network predicted too high growth rates, while
the ME-matrix growth rates were often below the experimentally measured ones (Figure Main Paper 2A). This is mainly caused by the parameters used for remaining amino acids and GAM in the biomass reaction of the ME-matrix. As the results of the sensitivity analysis showed (Figure 2), these two parameters play a key role in accurate growth rate prediction and will require adjustment to match the measured growth rates.

Comparison of Biolog data

Biolog data (http://www.biolog.com/) were used to compare with predicted growth phenotypes of the ME-matrix. Overall, the ME-matrix predicted 128 of 170 growth phenotypes correctly (75%). Moreover, the ME-matrix showed improved prediction in 14 cases compared to iAF1260 but worsen the prediction in 11 cases. In particular, the ME-matrix was able to use all 51 tested nitrogen sources for growth. From the tested carbon sources, 48 of the 87 supported growth in silico and in vivo. While it is valuable to know the number of correct growth phenotypes, the analysis of the false negative (FN) and false positive (FP) results is more interesting as this may lead to new biological discoveries. FPs (e.g., model can grow while no growth was observed experimentally) hint towards missing regulation. We identified 32 FP cases, where half of them were on carbon sources and the other half on nitrogen sources. The FP growth on nitrogen may also be caused by the fact, that the carbon source is not known, which was used in the Biolog data. Interestingly, the ME-matrix corrected six cases of FP predictions compared to iAF1260, which illustrates the further confined ME-matrix solution space. In contrast, FN predictions indicate missing links in the network. Since no link was removed from iAF1260, growth conditions, which did not support growth of the ME-matrix, but of iAF120, were caused by the additional constraints (e.g., stoichiometric synthesis constraints or coupling constraints). For instance, four carbon sources did not supported ME-matrix growth. One of these carbon sources is formate, which showed weak growth in vivo and in iAF1260 (Feist et al., 2007). We tested growth of ME-matrix at various formate uptake rates but no growth could be observed. Two of the other FNs were glucose-1-phosphate and fructose-6-phosphate, which did not supported growth in the previous E. coli metabolic reconstruction. Further analysis will be required to elucidate why the ME-matrix is not able to grow on these two media. Taken together, our results show that the growth phenotype of the ME-matrix is comparable with the metabolic reconstruction of E. coli. This result was somehow expected as the metabolic reconstruction served as baseline for the ME-matrix.

Gene deletion analysis

We used the ME-matrix to determine in silico growth phenotypes for single gene knockout strains in glycerol minimal medium. We considered only the 1,260 metabolic genes and compared the predictions with in vivo essential genes (Joyce and Palsson, 2006; Baba et al., 2006). In 89% of the cases, our predictions agreed with the experimental data and iAF1260 predictions (Figure 3). More interestingly, the ME-matrix improved prediction of six essential genes (Figure 3), which were non-essential in silico when the metabolic network (iAF1260) was used alone (Joyce and Palsson, 2006; Feist et al., 2007).

Reduced Cost

Reduced cost (RC) is a parameter of linear programming (LP) problems, which is associated with each network reaction ($v_i$) and represents the amount by which the objective function (e.g., growth rate) could be increased when the flux rate through this reaction was increased by a single unit (Ramakrishna et al., 2001). RC is often used to analyze the obtained optimal solution and evaluate alternate solutions from the original solution. In this study, we use the RC analysis to identify constraining reaction rates in the model. Therefore, we analyzed the RC of the four simulated conditions (only wildtype data were used) (Figure Main Paper 2A). We found that the transcription initiation reactions of the rRNA operons had the greatest reduced cost associated in all four conditions. This result was somehow expected, as ribosome synthesis rate and biomass production are competing for resources.

In silico strain collection

The 15 in silico CUB mutant strains were generated as described in the Material & Method section. The genetic code providing the basis for the ME-matrix and the modeled tRNA reading are listed in Figure 4 and
Tables 3, 4. tRNA species that can read multiple codons are listed for each codon. If a codon is recognized by multiple tRNA species, a generic tRNA species has been added to the model along with reactions converting the reading tRNA species into the generic one (see also (Thiele et al., 2009) for details). This formulation permits to model the complexity of tRNA reading while not needing to write alternate translation reactions. The use of generic tRNA species also highlights the redundancy in the codon reading.

**Codon Usage**

It can be assumed that many of the genes in the ME-matrix are highly transcribed and translated in the cell due to their central function, i.e., being involved in macromolecular synthesis machinery or central metabolism. However, we find that the codon usage of wildtype ME-matrix is comparable with genome codon usage (Figure 5).

**Entropy**

One approach to assess the extend of changes to the gene sequence that we introduced by biasing or equilibrating the codon usage is via an entropy measure, which reflects how biased or unbiased the codon usage of the 1,823 genes in the strains is compared to a random distribution. As expected, the equilibrated strains had the highest entropy (thus having most random sequence), while the biased strains had lower entropy than the wildtype strain (Figure 6). No obvious correlation between entropy value and maximal achievable growth rate could be observed, except that high entropy seem to have a reducing effect on the growth rates.
Supplementary Figures

Figure 1: **Schematic depiction of the network integration.** The metabolic network (M-matrix) and transcriptional/translation network (E-matrix) were into a combined, integrated network (ME-matrix) with changed steady-state solution space.
Figure 2: Sensitivity analysis. We tested the sensitivity of the predicted growth rate as a function of the remaining amino acid (AA) requirement in the biomass function and as a function of the remaining growth associated maintenance (GAM) that is left in the biomass function. The experimentally observed growth rate is shown with the dotted line. Since the ME-matrix covers about 1,900 of 4,400 E. coli genes, we decided to allocate 50% of the AA requirements and the 50% of the GAM for the ME-matrix genes and gene products. This plot also highlights that finetuning of these two parameters will be important to obtain accurate predictions in growth rate.

![Sensitivity analysis graph]

Figure 3: Gene essentiality. Comparison of gene essentiality in the ME-matrix and the metabolic network used in Joyce et al. Joyce et al. (2006)
Figure 4: **Degeneracy of the genetic code.** A. Genetic code employed in this study. Number of cognate tRNAs per amino acid is given in parenthesis. B. Schematic illustration of the degeneracy of genetic code.

86 tRNAs
60 amino acid coding codons
20 amino acids
1 start codon
3 stop codons

Figure 5: **Codon usage.** Comparison of codon usage in ME-matrix associated genes and across the genome.
Figure 6: **Genome entropy versus maximal possible growth rate.** The 16 *in silico* strains are shown with their predicted growth rates in glucose minimal medium/aerobic condition (GlcAer), glucose minimal medium/anaerobic condition (GlcAnaer), glycerol minimal medium/aerobic condition (Glyc) and lactate minimal medium, aerobic conditions (Lac). Eq. strains = equilibrated strains. WT = wildtype.
Supplementary Tables

Table 1: Metallo-ions and prosthetic groups included in the ME-matrix.

<table>
<thead>
<tr>
<th>Name</th>
<th>Abbr.</th>
<th># of proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Metallo-ions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calcium</td>
<td>Ca2</td>
<td>4</td>
</tr>
<tr>
<td>Cobalt</td>
<td>cobalt2</td>
<td>7</td>
</tr>
<tr>
<td>Copper</td>
<td>Cu2</td>
<td>2</td>
</tr>
<tr>
<td>Iron(II)</td>
<td>Fe2</td>
<td>13</td>
</tr>
<tr>
<td>iron-sulfur-cluster</td>
<td>$[Fe_2S_2]^{2+}$</td>
<td>34</td>
</tr>
<tr>
<td>iron-sulfur-cluster</td>
<td>$[Fe_4S_4]^{2+}$</td>
<td>2</td>
</tr>
<tr>
<td>Kalium</td>
<td>K</td>
<td>12</td>
</tr>
<tr>
<td>Magnesium</td>
<td>Mg2</td>
<td>173</td>
</tr>
<tr>
<td>Manganese</td>
<td>Mn2</td>
<td>27</td>
</tr>
<tr>
<td>Molybdate</td>
<td>Mo</td>
<td>1</td>
</tr>
<tr>
<td>Nickel</td>
<td>Ni2</td>
<td>2</td>
</tr>
<tr>
<td>Zinc</td>
<td>Zn2</td>
<td>44</td>
</tr>
<tr>
<td><strong>Prosthetic groups</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2′-(5′-triphosphoribosyl)-3′-dephospho-CoA</td>
<td>2tpr3dpcoa</td>
<td>1</td>
</tr>
<tr>
<td>Adenosylcobalamin</td>
<td>adocbl</td>
<td>1</td>
</tr>
<tr>
<td>Biotin</td>
<td>btn</td>
<td>1</td>
</tr>
<tr>
<td>Flavin adenine dinucleotide (oxidized)</td>
<td>fad</td>
<td>21</td>
</tr>
<tr>
<td>Flavomononucleotide (oxidized)</td>
<td>fmn</td>
<td>9</td>
</tr>
<tr>
<td>Flavomononucleotide (reduced)</td>
<td>fmnh2</td>
<td>1</td>
</tr>
<tr>
<td>Nicotinamide adenine dinucleotide</td>
<td>nad</td>
<td>6</td>
</tr>
<tr>
<td>Nicotinamide adenine dinucleotide phosphate</td>
<td>nadp</td>
<td>2</td>
</tr>
<tr>
<td>Pyridoxal 5’-phosphate</td>
<td>pydx5p</td>
<td>42</td>
</tr>
<tr>
<td>Protoheme</td>
<td>pheme</td>
<td>6</td>
</tr>
<tr>
<td>Siroheme</td>
<td>sheme</td>
<td>2</td>
</tr>
<tr>
<td>Thiamine diphosphate</td>
<td>thmpp</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 2: Information used for synthesis reactions of *E. coli*’s metabolic genes.

<table>
<thead>
<tr>
<th>Information</th>
<th>Subsystem/Reaction</th>
<th>Source/Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription unit</td>
<td>Transcription</td>
<td>EcoCyc (Karp et al., 2004)</td>
</tr>
<tr>
<td>Gene coordinate, direction</td>
<td>Transcription</td>
<td>Riley <em>et al.</em> (Riley et al., 2006)</td>
</tr>
<tr>
<td>Gene function</td>
<td>Metabolism</td>
<td>iAF1260 (Feist et al., 2007)</td>
</tr>
<tr>
<td>Protein information</td>
<td>Protein complex formation</td>
<td>iAF1260 (Feist et al., 2007), EcoCyc (Karp et al., 2004), primary literature</td>
</tr>
<tr>
<td>Metallo-ion</td>
<td>Protein maturation</td>
<td>EcoCyc (Karp et al., 2004), protein structure, primary literature</td>
</tr>
<tr>
<td>Prosthetic group</td>
<td>Protein maturation</td>
<td>EcoCyc (Karp et al., 2004), protein structure, primary literature</td>
</tr>
<tr>
<td>Codons</td>
<td>tRNA</td>
<td>Codon</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------</td>
<td>-------</td>
</tr>
<tr>
<td>alaT, alaU, alaV</td>
<td>ala1-tRNA</td>
<td>gct</td>
</tr>
<tr>
<td>alaT, alaU, alaV</td>
<td>ala1-tRNA</td>
<td>gca</td>
</tr>
<tr>
<td>alaT, alaU, alaV</td>
<td>ala1-tRNA</td>
<td>gcg</td>
</tr>
<tr>
<td>alaW, alaX</td>
<td>ala2-tRNA</td>
<td>gcc</td>
</tr>
<tr>
<td>argQ, argV, argY, argZ</td>
<td>arg1-tRNA</td>
<td>cgt</td>
</tr>
<tr>
<td>argQ, argV, argY, argZ</td>
<td>arg1-tRNA</td>
<td>cgc</td>
</tr>
<tr>
<td>argQ, argV, argY, argZ</td>
<td>arg1-tRNA</td>
<td>cga</td>
</tr>
<tr>
<td>argU-tRNA</td>
<td>argW-tRNA</td>
<td>agg</td>
</tr>
<tr>
<td>argX-tRNA</td>
<td>argX-tRNA</td>
<td>cgg</td>
</tr>
<tr>
<td>asnT, asnU, asnV, ansW</td>
<td>asn1-tRNA</td>
<td>aac</td>
</tr>
<tr>
<td>asnT, asnU, asnV, ansW</td>
<td>asn1-tRNA</td>
<td>aat</td>
</tr>
<tr>
<td>aspT, aspU, aspV</td>
<td>asp1-tRNA</td>
<td>gac</td>
</tr>
<tr>
<td>aspT, aspU, aspV</td>
<td>asp1-tRNA</td>
<td>gat</td>
</tr>
<tr>
<td>cysT-tRNA</td>
<td>cysT-tRNA</td>
<td>tgc</td>
</tr>
<tr>
<td>gltT, gltU, gltV, gltW</td>
<td>glu1-tRNA</td>
<td>gaa</td>
</tr>
<tr>
<td>gltT, gltU, gltV, gltW</td>
<td>glu1-tRNA</td>
<td>gag</td>
</tr>
<tr>
<td>glyU-tRNA</td>
<td>glyU-tRNA</td>
<td>gga</td>
</tr>
<tr>
<td>glyU-tRNA</td>
<td>glyU-tRNA</td>
<td>ggg</td>
</tr>
<tr>
<td>hisR-tRNA</td>
<td>hisR-tRNA</td>
<td>cac</td>
</tr>
<tr>
<td>hisR-tRNA</td>
<td>hisR-tRNA</td>
<td>cat</td>
</tr>
<tr>
<td>ileT, ileU, ileV</td>
<td>ile1-tRNA</td>
<td>atc</td>
</tr>
<tr>
<td>ileT, ileU, ileV</td>
<td>ile1-tRNA</td>
<td>att</td>
</tr>
<tr>
<td>ileX, ileY</td>
<td>ile2-tRNA</td>
<td>ata</td>
</tr>
<tr>
<td>leuP, leuQ, leuT, leuV, leuW</td>
<td>leu1-tRNA</td>
<td>ctc</td>
</tr>
<tr>
<td>leuP, leuQ, leuT, leuV, leuW</td>
<td>leuU-tRNA</td>
<td>ctc</td>
</tr>
<tr>
<td>Codons recognition by tRNA in the ME-matrix. Part II.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>generic tRNA</strong></td>
<td><strong>codon</strong></td>
<td><strong>Amino acid</strong></td>
</tr>
<tr>
<td>leuU-tRNA</td>
<td>ctt</td>
<td>leu-L</td>
</tr>
<tr>
<td>leuW-tRNA</td>
<td>cta</td>
<td>leu-L</td>
</tr>
<tr>
<td>leu2-tRNA</td>
<td>ttg</td>
<td>leu-L</td>
</tr>
<tr>
<td>leuZ-tRNA</td>
<td>tta</td>
<td>leu-L</td>
</tr>
<tr>
<td>lysQ, lysT, lysV, lysW, lysY, lysZ</td>
<td>lys1-tRNA</td>
<td>aza</td>
</tr>
<tr>
<td>lysQ, lysT, lysV, lysW, lysY, lysZ</td>
<td>lys1-tRNA</td>
<td>aag</td>
</tr>
<tr>
<td>metT, metU</td>
<td>met1-tRNA</td>
<td>atg</td>
</tr>
<tr>
<td>pheU, pheV</td>
<td>phe1-tRNA</td>
<td>ttc</td>
</tr>
<tr>
<td>pheU, pheV</td>
<td>phe1-tRNA</td>
<td>ttt</td>
</tr>
<tr>
<td>proK, proM</td>
<td>pro1-tRNA</td>
<td>ccc</td>
</tr>
<tr>
<td>proL, proM</td>
<td>pro2-tRNA</td>
<td>cct</td>
</tr>
<tr>
<td>serW, serX</td>
<td>ser1-tRNA</td>
<td>tcc</td>
</tr>
<tr>
<td>ser2-tRNA</td>
<td>tct</td>
<td>ser-L</td>
</tr>
<tr>
<td>ser3-tRNA</td>
<td>tca</td>
<td>ser-L</td>
</tr>
<tr>
<td>serT, serU</td>
<td>serV-tRNA</td>
<td>agc</td>
</tr>
<tr>
<td>serV-tRNA</td>
<td>agt</td>
<td>ser-L</td>
</tr>
<tr>
<td>thrT, thrV</td>
<td>thr1-tRNA</td>
<td>acc</td>
</tr>
<tr>
<td>thr2-tRNA</td>
<td>act</td>
<td>thr-L</td>
</tr>
<tr>
<td>thr3-tRNA</td>
<td>aec</td>
<td>thr-L</td>
</tr>
<tr>
<td>tyrT, tyrU, tyrV</td>
<td>tyr1-tRNA</td>
<td>tac</td>
</tr>
<tr>
<td>tyr1-tRNA</td>
<td>tat</td>
<td>tyr-L</td>
</tr>
<tr>
<td>valT, valU, valX, valY, valZ</td>
<td>val1-tRNA</td>
<td>gta</td>
</tr>
<tr>
<td>val1-tRNA</td>
<td>gtc</td>
<td>val-L</td>
</tr>
<tr>
<td>val2-tRNA</td>
<td>gtc</td>
<td>val-L</td>
</tr>
<tr>
<td>val3-tRNA</td>
<td>gtt</td>
<td>val-L</td>
</tr>
<tr>
<td>metV, metW, metY, metZ</td>
<td>fmet-tRNA</td>
<td>atg</td>
</tr>
</tbody>
</table>

### References


coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Molecular Systems Biology 3.


