J. theor. Biol. (2003) **220,** 1–18
doi:10.1006/jtbi.2003.3087, available online at http://www.idealibrary.com on IDEAL®

Sequence-Based Analysis of Metabolic Demands for Protein Synthesis in Prokaryotes

TIMOTHY E. ALLEN† AND BERNHARD Ø. PALSSON*†

†*Department of Bioengineering, University of California, San Diego, La Jolla, CA* 92093-0412, *U.S.A.*

(*Received on* 8 *February* 2002, *Accepted in revised form on* 15 *May* 2002)

Constraints-based models for microbial metabolism can currently be constructed on a genome-scale. These models do not account for RNA and protein synthesis. A scalable formalism to describe translation and transcription that can be integrated with the existing metabolic models is thus needed. Here, we developed such a formalism. The fundamental protein synthesis network described by this formalism was analysed via extreme pathway and flux balance analyses. The protein synthesis network exhibited one extreme pathway per messenger RNA synthesized and one extreme pathway per protein synthesized. The key parameters in this network included promoter strengths, messenger RNA half-lives, and the availability of nucleotide triphosphates, amino acids, RNA polymerase, and active ribosomes. Given these parameters, we were able to calculate a cell's material and energy expenditures for protein synthesis using a flux balance approach. The framework provided herein can subsequently be integrated with genome-scale metabolic models, providing a sequence-based accounting of the metabolic demands resulting from RNA and protein polymerization.

Introduction

The large number of genome sequences completed in recent years has underscored the need to develop genome-scale models that can be used to elucidate phenotypic behavior from the genotype (Edwards & Palsson, 1998; Schilling *et al.*, 1999). The available annotated sequences, along with known organism-specific biochemical and physiological data, have been implemented in the reconstruction of genome-scale models of metabolism (Karp *et al.*, 1996; Selkov *et al.*, 1998; Ogata *et al.*, 1999; Overbeek *et al.*, 2000; Covert *et al.*, 2001a).

Kinetic models are very difficult to construct on a genome-scale due to the sheer number of parameters required (Bailey, 2001). A constraints-based approach can be used to successfully circumvent this problem under certain conditions. Such an approach relies upon the fact that metabolic networks are constrained by physicochemical laws which limit what phenotypes the cell is capable of attaining (Palsson, 2000). Thus, rather than calculating a unique phenotypic solution, one can determine the closed solution space within which the steady-state solution must lie, thereby defining the metabolic capabilities of the cellular network. Linear programming (Chvátal, 1983) can then be used to determine the solution within this space that optimizes a specified cellular objective. This approach, called flux balance analysis (FBA) (Varma & Palsson, 1994; Bonarius *et al.*, 1997; Edwards *et al.*, 1999; Gombert & Nielsen, 2000), has been successfully applied to genome-scale

*Corresponding author. Tel.: +1-858-822-3120; fax: +1-858-822-3120.
E-mail address: palsson@ucsd.edu (B. Ø. Palsson).

0022-5193/03/$35.00

metabolic models of *Haemophilus influenzae* (Edwards & Palsson, 1999), *Escherichia coli* (Edwards & Palsson, 2000; Edwards *et al.*, 2001), *Helicobacter pylori* (Schilling *et al.*, in press), and *Saccharomyces cerevisiae* (Famili *et al.*, in review; Förster *et al.*, in review).

Existing constraints-based genome-scale metabolic networks do not include sequence-based macromolecular polymerization reactions—namely, RNA and protein synthesis—except lumped as monomeric amino acid and nucleotide triphosphate demands for cellular growth (Varma & Palsson, 1993). These monomeric demands are determined from the cellular biomass constituents (Neidhardt *et al.*, 1990) and are thus independent of genome sequence. There is consequently a need to develop a constraints-based formalism for RNA and protein synthesis. Furthermore, this formalism needs to be readily scalable to the genome-scale.

General models of protein synthesis have included non-sequence-dependent models within genome-scale metabolic networks (Tomita *et al.*, 1999) and mechanistically detailed kinetic, but not genome-scale models (Peretti & Bailey, 1986; Drew, 2001). Detailed kinetic models have been developed for individual genes and operons and the proteins for which they encode, including the *lac* operon (Wong *et al.*, 1997) and the *trp* operon (Sinha, 1988; Santillán & Mackey, 2001) in *E. coli*.

A sequence-based genome-scale model of protein synthesis, however, has not been developed, and currently no framework has been established for such a large-scale incorporation of protein synthesis to the current models. This paper describes a fundamental reaction scheme for protein synthesis that provides such a scalable framework. We analyse this basic network using flux balance and extreme pathway analyses, and we identify the parameters that govern both gene expression and protein synthesis.

Methods

FUNDAMENTAL REACTION SCHEME FOR PROTEIN PRODUCTION

In order to develop a scalable framework within which to describe protein synthesis, it is necessary to identify the fundamental reactions that comprise an "idealized protein production scheme" (Fig. 1). These reactions will comprise an "elemental system" for a particular gene and its protein, having conceptual systemic boundaries across which the building blocks and energy metabolites for polynucleotide and protein polymerization will be exchanged.

For a given gene, G, and the protein for which it encodes, we can write such a fundamental set of reactions (Table 1). This fundamental reaction set contains one gene encoding for one protein, and one type each of nucleotide, amino acid, and transfer RNA (tRNA), and is illustrated in Fig. 1. The first six fluxes in Table 1 correspond to fluxes internal to the system, and the last nine correspond to exchange fluxes. A summary of all abbreviations and symbols used is provided in Appendix A. The reaction set is as follows:

- *Transcription initiation*: The reaction corresponding to the flux, v_1, describes the binding of RNA polymerase (RNAP) to the promoter of G to form the open-promoter complex, G^*. This reaction is usually referred to as transcription initiation. We assume that the forward reaction from the closed RNAP–promoter complex to the open complex (G^*) is much faster than the reverse reaction (Record *et al.*, 1996); thus, v_1 is essentially irreversible.
- *Transcription elongation*: The RNAP then proceeds along the gene during elongation, incorporating nucleotide triphosphates (NTPs) in a series of polymerization reactions represented by v_2. For every NTP added, a pyrophosphate (PP_i) will be released. The liberated PP_i will immediately be hydrolysed by pyrophosphatase into two inorganic phosphates (P_i) to drive the reaction (v_2).
- *mRNA degradation*: The messenger RNA (mRNA) that is produced by v_2 will subsequently be degraded into its nucleotide monophosphate (NMP) constituents (v_3).
- *Translation initiation*: The mRNA will also bind to free ribosomes (translation initiation) to form the active ribosomal complex, rib* (v_4).
- *Translation elongation*: The polymerization reactions that incorporate amino acids (AAs) in the synthesis of the complete protein are lumped into v_5. Two GTPs are required per AA

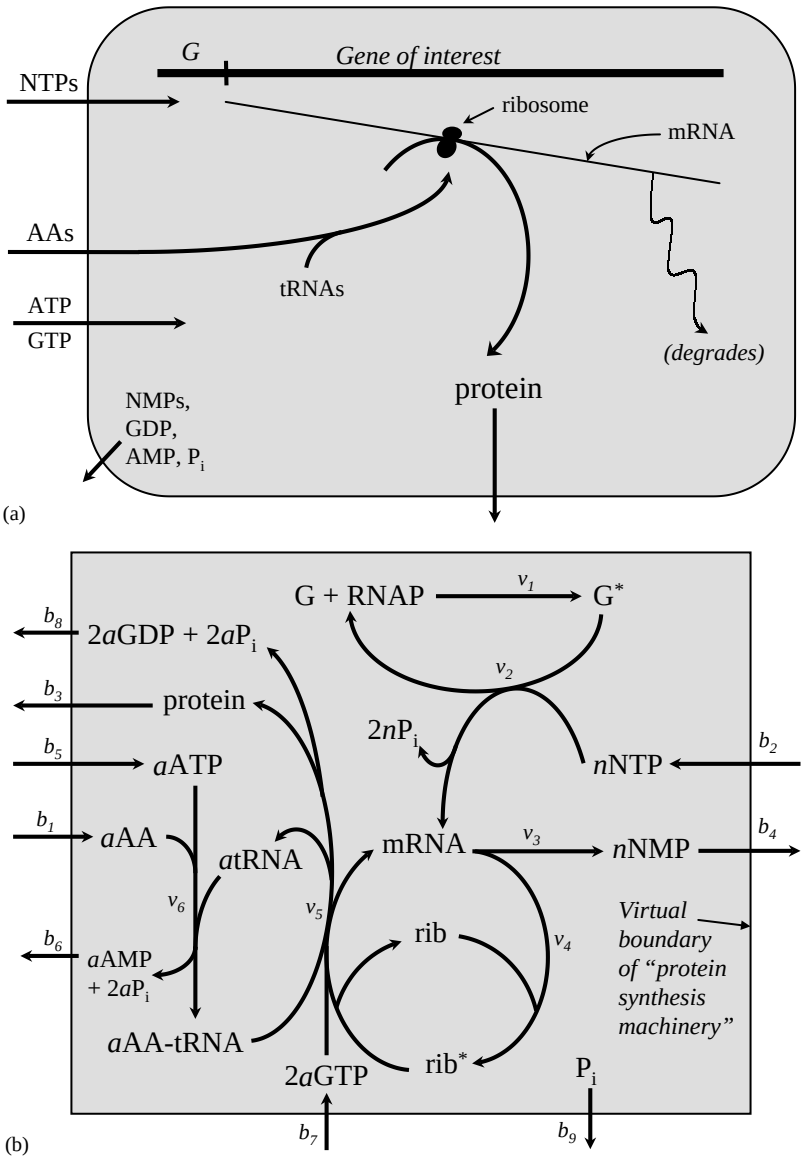

FIG. 1. The fundamental reaction scheme for protein synthesis. Panel (a) provides a simplified schematic for the synthesis of a protein encoded by a generic gene. Panel (b) gives the complete fundamental reaction network based upon the individual reactions listed in Table 1 and discussed in the text.

incorporated: one in the binding of the charged transfer RNA (AA–tRNA) to the ribosomal A-site, and the other in the translocation of the amino acid (with the rest of the nascent polypeptide) from the A- to the P-site.

• *tRNA charging*: In order to recharge the tRNAs, each AA binds ATP to form aminoacyl-AMP and PP_i. The aminoacyl-AMP then reacts with a tRNA to produce the AA–tRNA and an AMP. These two reactions, driven by the hydrolysis of PP_i to $2P_i$ by pyrophosphatase, are represented by v_6.

The AA and NTP inputs represent the building blocks (b_1, b_2), and the ATP and GTP inputs (b_5, b_7) represent the energy cost for the production of protein; the NMP, AMP, GDP, and P_i outputs represent the by-products (b_4, b_6, b_8, b_9); and the protein output (b_3) is simply the production rate of the protein.

This fundamental reaction set applies for any gene, G, regardless of the number of nucleotides or amino acids incorporated. For those prokaryotic genes which are present in operons, the entire operon is transcribed into a single mRNA.

TABLE 1
Simplified, fundamental reaction set for protein production

Transcription initiation:	$G + \text{RNAP} \xrightarrow{v_1} G^*$
Transcription:	$G^* + n\text{NTP} \xrightarrow{v_2} \text{mRNA} + G + \text{RNAP} + 2n\text{P}_i$
mRNA decay:	$\text{mRNA} \xrightarrow{v_3} n\text{NMP}$
Translation initiation:	$\text{mRNA} + \text{rib} \xrightarrow{v_4} \text{rib}^*$
Translation:	$\text{rib}^* + a\text{AAtRNA} + 2a\text{GTP} \xrightarrow{v_5} a\text{tRNA} + 2a\text{GDP} + 2a\text{P}_i + \text{rib} + \text{mRNA} + \text{protein}$
tRNA charging:	$\text{AA} + \text{tRNA} + \text{ATP} \xrightarrow{v_6} \text{AMP} + 2\text{P}_i + \text{AAtRNA}$
Exchange fluxes:	$\text{AA}_{ext} \xrightarrow{b_1} \text{AA}$
	$\text{NTP}_{ext} \xrightarrow{b_2} \text{NTP}$
	$\text{protein} \xrightarrow{b_3} \text{protein}_{ext}$
	$\text{NMP} \xrightarrow{b_4} \text{NMP}_{ext}$
	$\text{ATP}_{ext} \xrightarrow{b_5} \text{ATP}$
	$\text{AMP} \xrightarrow{b_6} \text{AMP}_{ext}$
	$\text{GTP}_{ext} \xrightarrow{b_7} \text{GTP}$
	$\text{GDP} \xrightarrow{b_8} \text{GDP}_{ext}$
	$\text{P}_i \xrightarrow{b_9} \text{P}_{i\,ext}$

Note: The first six reactions, which are discussed in the text, occur within the virtual systemic boundary within which the machinery for protein synthesis resides. The last nine reactions correspond to the exchange of building blocks (i.e. AAs and NTPs), protein, by-products (e.g. NMPs), and energy molecules (i.e. ATP and GTP) across the systemic boundary.

This polycistronic mRNA is then involved in separate reactions for each of the proteins for which it encodes. Thus, a set of reactions $v_{1,2,3}$ is written for every mRNA being produced, and reactions $v_{4,5}$ are written for every protein being synthesized.

THE PRODUCTION OF COMPONENTS OF THE PROTEIN SYNTHESIS "MACHINERY"

The internal (to the synthesis system defined) production of ribosomes, transfer RNA, and RNA polymerase to the fundamental reaction scheme discussed above will add the reactions listed in Table 2 to the fundamental reaction set (Fig. 2). For the untranslated RNA transcripts (i.e. tRNA and ribosomal RNA), production fluxes analogous to v_1 and v_2 are written. A degradation flux is not added for the untranslated transcripts since they are many times more stable than mRNA transcripts and thus unlikely to degrade within the time-scale of cellular growth (Nierlich, 1978). The synthesis of RNAP is analogous to that of the generic protein in the fundamental system, except that no exchange flux is included for RNAP since it remains internal to the system (Fig. 2).

FLUX BALANCE ANALYSIS (FBA)

FBA has been reviewed in detail previously (Varma & Palsson, 1994; Bonarius *et al.*, 1997; Edwards *et al.*, 1999; Gombert & Nielsen, 2000). In short, a mass balance can be described for a system (e.g. the components of a cell's protein production machinery), which in a steady state can be written as

$$\mathbf{Sv} = \mathbf{0}, \tag{1}$$

where $\mathbf{S}$ is the $m \times n$ stoichiometric matrix (having m metabolites and n reaction fluxes; e.g. see Table 3) and $\mathbf{v}$ is an $n \times 1$ vector containing the values of the fluxes through the reactions involved in the system. These fluxes will be subject to thermodynamic and capacity constraints (e.g. v_{max}'s of promoter bindings, maximum elongation rates, etc., as discussed below), described in general by inequality constraints of the form

$$\alpha_i \leqslant v_i \leqslant \beta_i, \tag{2}$$

TABLE 2
Reactions added to the fundamental system when including the internal production of RNAP, tRNA, and rRNA

RNAP:	$G_P + \text{RNAP} \overset{v_{1P}}{\rightarrow} G_P^*$	
	$G_P^* + n_P\text{NTP} \overset{v_{2P}}{\rightarrow} \text{mRNA}_P + G_P + \text{RNAP} + 2n_P\text{P}_i$	
	$\text{mRNA}_P \overset{v_{3P}}{\rightarrow} n_P\text{NMP}$	
	$\text{mRNA}_P + \text{rib} \overset{v_{4P}}{\rightarrow} \text{rib}_P^*$	
	$\text{rib}_P^* + a_P\text{AAtRNA} + 2a_P\text{GTP} \overset{v_{5P}}{\rightarrow} a_P\text{tRNA} + 2a_P\text{GDP} + 2a_P\text{P}_i + \text{rib} + \text{mRNA}_P + \text{RNAP}$	
tRNA:	$G_t + \text{RNAP} \overset{v_{1t}}{\rightarrow} G_t^*$	
	$G_t^* + n_t\text{NTP} \overset{v_{2t}}{\rightarrow} \text{tRNA} + G_t + \text{RNAP} + 2n_t\text{P}_i$	
rRNA:	$G_r + \text{RNAP} \overset{v_{1r}}{\rightarrow} G_r^*$	
	$G_r^* + n_r\text{NTP} \overset{v_{2r}}{\rightarrow} \text{rib} + G_r + \text{RNAP} + 2n_r\text{P}_i$	

Note: The subscript P denotes RNAP, the subscript t denotes tRNA, and the subscript r denotes rRNA. For example, G_t refers to the gene encoding for tRNA, and rib_P^* refers to the actively translating ribosomal complex for the synthesis of RNAP.

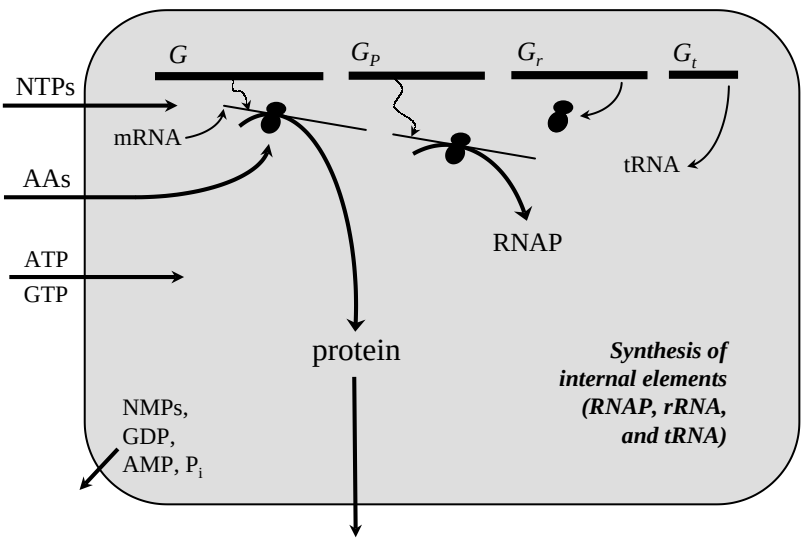

FIG. 2. Addition of internal accessory elements to the fundamental protein synthesis scheme depicted in Fig. 1. The basic reaction scheme is provided, in addition to the internal production of the protein synthesis machinery (RNAP, tRNA, and ribosomes).

where α_i and β_i represent the lower and upper bounds constraining each flux. Linear programming can be used to maximize protein production (i.e. b_3), given the stated constraints. Protein production is chosen as the objective in this study to determine, for a given set of environmental conditions and resources, how much the protein synthesis machinery within the cell can produce. Optimal flux distributions in this study were identified using a commercially available linear programming package (LINDO, Lindo Systems, Chicago), subject to the constraints given in eqns (1) and (2).

EXTREME PATHWAY ANALYSIS

Since we have assumed that all of the reactions in Table 1 are essentially irreversible (i.e. $v_i \geqslant 0$),

TABLE 3

The stoichiometric matrix for the fundamental reaction scheme given in Table 2 and depicted in Fig. 2

$S =$

v_1	v_{1P}	v_{1t}	v_{1r}	v_2	v_{2P}	v_{2t}	v_{2r}	v_3	v_{3P}	v_4	v_{4P}	v_5	v_{5P}	v_6	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	
−1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	G
0	−1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	G_{RNAP}
0	0	−1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	G_{tRNA}
0	0	0	−1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	G_{rRNA}
−1	−1	−1	−1	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	RNAP
1	0	0	0	−1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	G^*
0	1	0	0	0	−1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	G^*_{RNAP}
0	0	1	0	0	0	−1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	G^*_{tRNA}
0	0	0	1	0	0	0	−1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	G^*_{rRNA}
0	0	0	0	$-n$	$-n_P$	$-n_t$	$-n_r$	0	0	0	0	0	0	0	0	−1	0	0	0	0	0	0	0	NTP
0	0	0	0	$2n$	$2n_P$	$2n_t$	$2n_r$	0	0	0	0	$2a$	$2a_P$	2	0	0	0	0	0	0	0	0	1	P_i
0	0	0	0	0	0	0	0	n	n_P	0	0	0	0	0	0	0	0	1	0	0	0	0	0	NMP
0	0	0	0	1	0	0	0	−1	0	−1	0	1	0	0	0	0	0	0	0	0	0	0	0	mRNA
0	0	0	0	0	1	0	0	0	−1	0	−1	0	1	0	0	0	0	0	0	0	0	0	0	mRNA_P
0	0	0	0	0	0	0	1	0	0	−1	−1	1	1	0	0	0	0	0	0	0	0	0	0	rib
0	0	0	0	0	0	1	0	0	0	0	0	a	a_P	−1	0	0	0	0	0	0	0	0	0	tRNA
0	0	0	0	0	0	0	0	0	0	1	0	−1	0	0	0	0	0	0	0	0	0	0	0	rib*
0	0	0	0	0	0	0	0	0	0	0	1	0	−1	0	0	0	0	0	0	0	0	0	0	rib^*_P
0	0	0	0	0	0	0	0	0	0	0	0	$-2a$	$-2a_P$	0	0	0	0	0	0	0	−1	0	0	GTP
0	0	0	0	0	0	0	0	0	0	0	0	$-a$	$-a_P$	1	0	0	0	0	0	0	0	0	0	AAtRNA
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	protein
0	0	0	0	0	0	0	0	0	0	0	0	$2a$	$2a_P$	0	0	0	0	0	0	0	0	1	0	GDP
0	0	0	0	0	0	0	0	0	0	0	0	0	0	−1	0	0	0	0	−1	0	0	0	0	ATP
0	0	0	0	0	0	0	0	0	0	0	0	0	0	−1	−1	0	0	0	0	0	0	0	0	AA
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	AMP

Note: The production of the protein synthesizing machinery is included.

extreme pathway analysis may be used to generate a unique set of vectors spanning the nullspace of **S** (Schilling *et al.*, 2000). A cone can be generated from this convex basis to circumscribe all allowable steady-state solutions to eqn (1):

$$C = \left\{ \mathbf{v} \,:\, \mathbf{v} = \sum_{i=1}^{k} \alpha_i \mathbf{p}_i,\ \alpha_i \geqslant 0,\ \forall i \right\}, \quad (3)$$

where $\mathbf{p}_i$ are the pathway vectors and α_i are positive weighting coefficients for each extreme pathway.

The pathway classification scheme developed previously (Schilling *et al.*, 2000) characterizes extreme pathways based on the activity of their exchange fluxes. Exchange fluxes can either be primary exchange fluxes (e.g. exchange of primary metabolites such as AAs, protein, etc.) or currency exchange fluxes (e.g. exchange of currency metabolites such as ATP, GTP, P_i, etc.). An extreme pathway is classified as:

- Type I if it contains any non-zero primary exchange flux;
- Type II if the only active exchange fluxes are for currency metabolites; or
- Type III if there are no active exchange fluxes in the extreme pathway.

Here, we use a variant on this classification scheme in that we consider the NTPs used in RNA polymerization reactions to be primary metabolites rather than currency metabolites, since they are building blocks for the RNA and not simply an energy supply. Thus, in the fundamental system described in Fig. 1, the primary exchange fluxes include b_1, b_2, b_3, and b_4.

FURTHER CONSTRAINTS UPON THE SYSTEM

Mass balance of mRNA production and degradation at steady state implies $v_1 = v_2 = v_3$ and $v_4 = v_5$. The first three fluxes, $v_{1,2,3}$, cannot be directly coupled to the fluxes $v_{4,5}$ except in the following fashion.

If a particular protein is synthesized, then $v_{4,5} > 0$. Then, if $v_{4,5} > 0$, it must also hold that $v_{1,2,3} > 0$, since some amount of transcript must be present (and thus be maintained) for translation to occur. This constraint can be written as

$$v_{1,2,3} = \min(v_{pr.str.},\ \text{limiting elongation rate}, b_2), \quad (4)$$

where $v_{pr.str.}$ is the transcription initiation flux which depends upon the promoter strength under the given set of conditions. The limiting elongation rate is the maximum speed at which the RNA polymerization can take place for a given transcript.

PROMOTER STRENGTHS

When constrained to non-zero values, the promoter-binding flux, v_1, can be set according to the promoter strength of the gene G under a specified set of conditions. Hence, if the influx of nucleotides is not limiting (i.e. $v_1 \leqslant nb_2$), the fluxes $v_{1,2,3}$ are set by the regulatory inputs to the system. If transcription regulation is to be taken into account, the regulatory "rules" under a specific set of conditions will determine whether or not v_1 is "on" (Covert *et al.*, 2001b), as well as its value under the specified conditions. For instance, if a gene is known to be down-regulated under a specific set of conditions, then the fluxes involved in synthesizing its mRNA transcript and the corresponding protein(s) will be set to zero. Bacterial promoter strengths have been studied extensively (Kajitani & Ishihama, 1983; Deuschle *et al.*, 1986; Zacharias *et al.*, 1991; Weller & Recknagel, 1994).

GLOBAL MAXIMUM ON TRANSCRIPTION ELONGATION

For particularly lengthy transcripts having strong promoters, it is possible that the elongation flux, v_2, will be limiting. Thus, an upper bound on the sum of all elongation (v_2) fluxes, $\beta_{2\text{-}global\ max}$, must be set based upon the sequence length and composition. For *E. coli*, typical elongation rates during transcription range from 50 to 100 nucleotides per RNAP s^{-1}, depending on the gene, the regulation present, and the growth rate (Gotta *et al.*, 1991; Vogel & Jensen, 1994). If we assume that there are approximately 2500–3000 RNAP molecules per cell (Iwakura *et al.*, 1974), then a typical cell is capable of

incorporating 125 000–300 000 nucleotides s^{-1} into RNA molecules. An average gene length of ~1000 bp (Blattner *et al.*, 1997) implies that 125–300 mRNA molecules can be made by a single cell every second, assuming that no rRNA and tRNA is being made. In reality, however, this estimate will be significantly less, since 80% of the total RNA in *E. coli* is rRNA, and 15% is tRNA, with only 4% comprised of mRNA (Neidhardt *et al.*, 1990).

CALCULATION OF MRNA CONCENTRATIONS

The mRNA degradation flux, v_3, is typically dependent upon the concentration of mRNA in a first-order, linear fashion:

$$v_3 = k[\text{mRNA}], \tag{5}$$

where k is the rate constant for the degradation of mRNA, which can be directly determined from the half-life of the particular mRNA (Iost & Dreyfus, 1995; Kushner, 1996). Once $v_{1,2,3}$ is determined from the promoter strength for G, from a limiting nucleotide influx, or from the global maximum on the elongation fluxes, we can calculate the concentration of mRNA as follows:

$$[\text{mRNA}] = \begin{cases} v_{1,2,3}/k & \text{if } v_{4,5} > 0, \\ 0 & \text{if } v_{4,5} = 0. \end{cases} \tag{6}$$

A Boolean logic representation has thus been assumed in that the gene is either "on" and being transcribed at a defined, condition-dependent rate, or it is "off," and is not being transcribed at all (Thomas, 1991). In reality, there is a very low basal level of transcription at all times; accordingly, stochastic models have been used to take into account the "leakiness" of all promoters (Arkin *et al.*, 1998). A small lower bound, α_1, on v_1 can thus be used. Hence, if the promoter strengths and mRNA half-lives for a given set of conditions are known (since both are subject to regulation), it is possible *a priori* to estimate genome-scale mRNA expression arrays.

GLOBAL MAXIMUM ON TRANSLATION INITIATION

Since there will be a finite number of free ribosomes available for protein synthesis at any given time, a global maximum must be set for the binding of each messenger RNA to a free ribosome (Laffend & Shuler, 1994; Fell, 2001). Thus, the following constraint must be applied to the v_4 fluxes for all of the proteins synthesized:

$$\sum_{i=1}^{N} v_{4,i} \leqslant \beta_{4\text{-}global\ max}, \tag{7}$$

where N is the total number of proteins being produced. Translation rates in *E. coli* are typically 16 amino acids per ribosome s^{-1} (or 48 nucleotides per ribosome s^{-1}) (Wagner, 2000), which corresponds to 299 200 amino acids per cell s^{-1}, assuming that there are 18 700 ribosomes per cell (Neidhardt *et al.*, 1990). Assuming an average open reading frame (ORF) length of 317 amino acids (Blattner *et al.*, 1997), a typical *E. coli* cell can produce ~950 protein molecules s^{-1}.

Results

The extreme pathway structure and key parameters governing protein production were identified for the following cases, ranging from highly simplified schemata to a whole bacterial operon:

1. Fundamental system having 1 gene, 1 type of nucleotide, and 1 type of amino acid (Fig. 1).
2. Fundamental system plus the production of the internal elements RNAP, tRNA, and rib (Fig. 2).
3. Generalized N-gene operon (polycistronic mRNA).
4. Fundamental system having biologically meaningful values of four types of nucleotides and 20 types of amino acids.
5. Production of malate dehydrogenase in *E. coli*.
6. Production of the proteins encoded by the *lac* operon in *E. coli*.

The extreme pathway results for the first four cases are summarized in Table 4.

CASE 1—FUNDAMENTAL SYSTEM

An extreme pathway analysis of the network considered in Case 1 divides the system into

TABLE 4
Fundamental extreme pathway structure for the first four cases of protein synthesis studied

EP	Net reaction equation	Primary function
Case 1		
p_1	$n\text{NTP} \rightarrow n\text{NMP} + 2n\text{P}_i$	mRNA maintenance
p_2	$a\text{AA} + a\text{ATP} + 2a\text{GTP} \rightarrow \text{protein} + a\text{AMP} + 2a\text{GDP} + 4a\text{P}_i$	mRNA utilization
Case 2		
p_1	$n_X\text{NTP} \rightarrow n_X\text{NMP} + 2n_X\text{P}_i$	prot. mRNA maint.
p_2	$n_Y\text{NTP} \rightarrow n_Y\text{NMP} + 2n_Y\text{P}_i$	RNAP mRNA maint.
p_3	$a\text{AA} + a\text{ATP} + 2a\text{GTP} \rightarrow \text{protein} + a\text{AMP} + 2a\text{GDP} + 4a\text{P}_i$	prot. mRNA util.
Case 3 (for a two-gene operon)		
p_1	$n\text{NTP} \rightarrow n\text{NMP} + 2n\text{P}_i$	mRNA maintenance
p_2	$a_A\text{AA} + a_A\text{ATP} + 2a_A\text{GTP} \rightarrow \text{protein}_A + a_A\text{AMP} + 2a_A\text{GDP} + 4a_A\text{P}_i$	mRNA utilization prod. of protein_A
p_3	$a_B\text{AA} + a_B\text{ATP} + 2a_B\text{GTP} \rightarrow \text{protein}_B + a_B\text{AMP} + 2a_B\text{GDP} + 4a_B\text{P}_i$	mRNA utilization prod. of protein_B
Case 4		
p_1	$n_1\text{N}_1\text{TP} + n_2\text{N}_2\text{TP} + n_3\text{N}_3\text{TP} + n_4\text{N}_4\text{TP} \rightarrow n_1\text{N}_1\text{MP} + n_2\text{N}_2\text{MP} + n_3\text{N}_3\text{MP} + n_4\text{N}_4\text{MP} + 2\sum_{i=1}^{4} n_i\text{P}_i$	mRNA maintenance
p_2	$a_1\text{AA}_1 + a_2\text{AA}_2 + \cdots + a_{20}\text{AA}_{20} + \sum_{i=1}^{20} a_i(\text{ATP} + 2\text{GTP}) \rightarrow \text{protein} + \sum_{i=1}^{20} a_i(\text{AMP} + 2\text{GDP} + 4\text{P}_i)$	mRNA utilization

Note: In Case 2, n_X denotes the number of NTPs in mRNA and n_Y denotes the number of NTPs in mRNA_P; in Case 3, the subscript A corresponds to the gene encoding protein A, and the subscript B corresponds to the gene encoding protein B; and in Case 4, the coefficients, $n_1, \ldots, n_4$ and $a_1, \ldots, a_{20}$ correspond to the number of each type of nucleotide/amino acid in the mRNA/protein of interest.

two functionally distinct categories (Fig. 3). One extreme pathway [Fig. 3(a)] corresponds to the maintenance of mRNA in the cell, and the other [Fig. 3(b)] corresponds to the utilization of mRNA to manufacture protein. The maintenance flux for a particular mRNA is required whenever the encoded protein is being produced.

If the gene, G, is being transcribed, one of the following two parameters limits the protein production flux (b_3): either the amino acid influx (b_1) or the maximum flux allowed for translation initiation due to the finite ribosomal pool (β_4), whichever is smaller. The limitation of the protein production flux for the fundamental system may be mathematically described as

$$b_3 \leqslant \min\left(\frac{b_1}{a}, \beta_4\right), \tag{8}$$

where a represents the number of amino acids in the protein. Note that, for this simplified system, there exists only one type of amino acid.

CASE 2—SYNTHESIS OF INTERNAL COMPONENTS

Untranslated RNA transcripts included in the system (i.e. rRNAs and tRNAs) will not result in any additional extreme pathways, since we have assumed that the stable RNA is allowed neither to degrade nor to leave this idealized system. When the RNAP gene is included, one extreme pathway is added for the maintenance of the associated mRNA, but no "production" pathway is added since no exchange flux is written for RNAP (Table 4; Fig. 4). The key parameters for protein production are thus unchanged from the system in Case 1. Thus, the internal elements are not included in any of the following cases.

CASE 3—GENERALIZED N-GENE OPERON

In the case of polycistronic mRNA, there will be an extreme pathway corresponding to the maintenance of the mRNA and one extreme pathway for each protein encoded for on that particular mRNA. In general, an N-gene operon encodes for N proteins, and the resulting system thus has $N + 1$ extreme pathways.

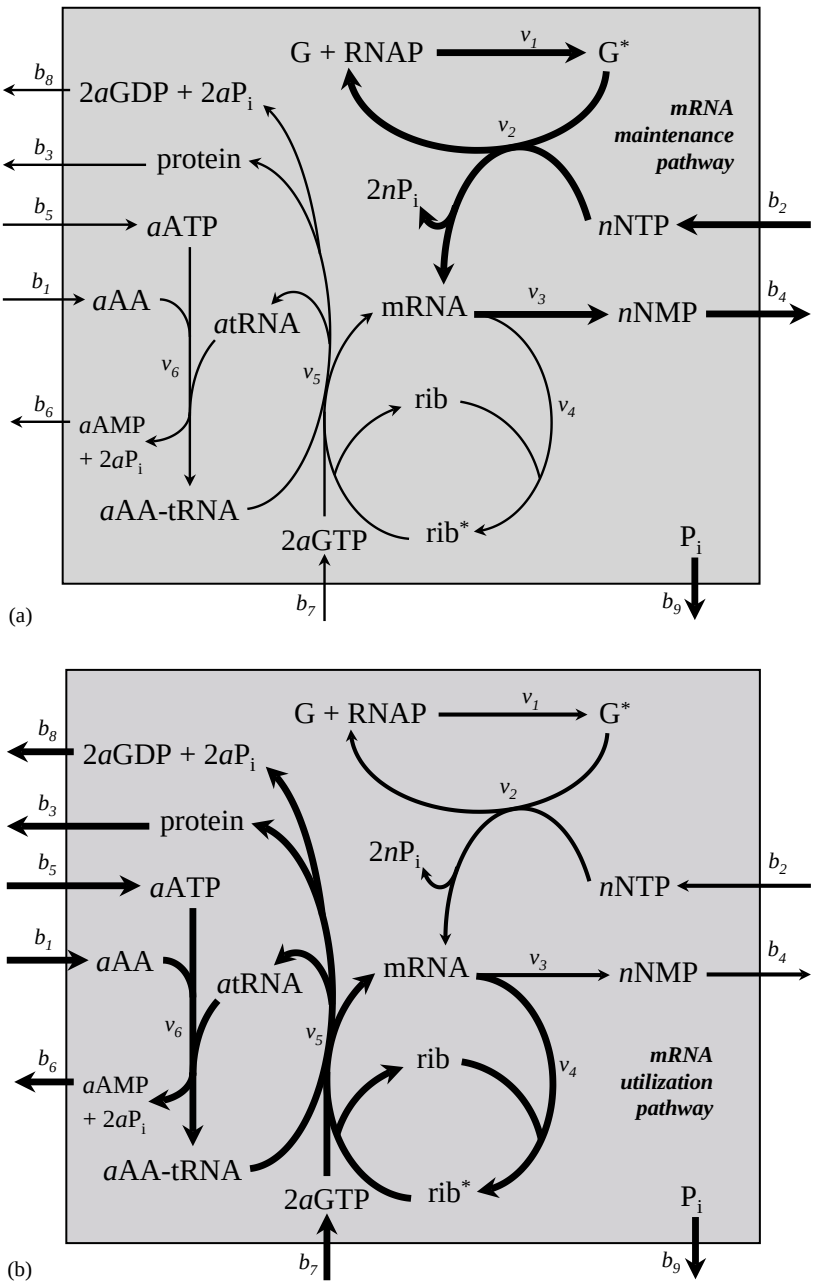

FIG. 3. The extreme pathway structure of the fundamental protein synthesis network. There are two extreme pathways for the basic system: (a) The extreme pathway for the maintenance of mRNA, and (b) the extreme pathway for the synthesis of protein.

The maximum production flux of each of the N proteins is determined in the following manner:

$$b_3 \leqslant \min\left(\frac{b_1}{\sum_{i=1}^{N} a_i}, \frac{\beta_{4\text{-}global\ max}}{N}\right). \qquad (9)$$

The ribosomal capacity is equally shared by the N transcripts since we assume that the half-lives are identical. (It is important to note that when the model is scaled to include more than one operon, the ribosomal capacity will be shared by all cellular transcripts and not just the transcripts of a particular operon).

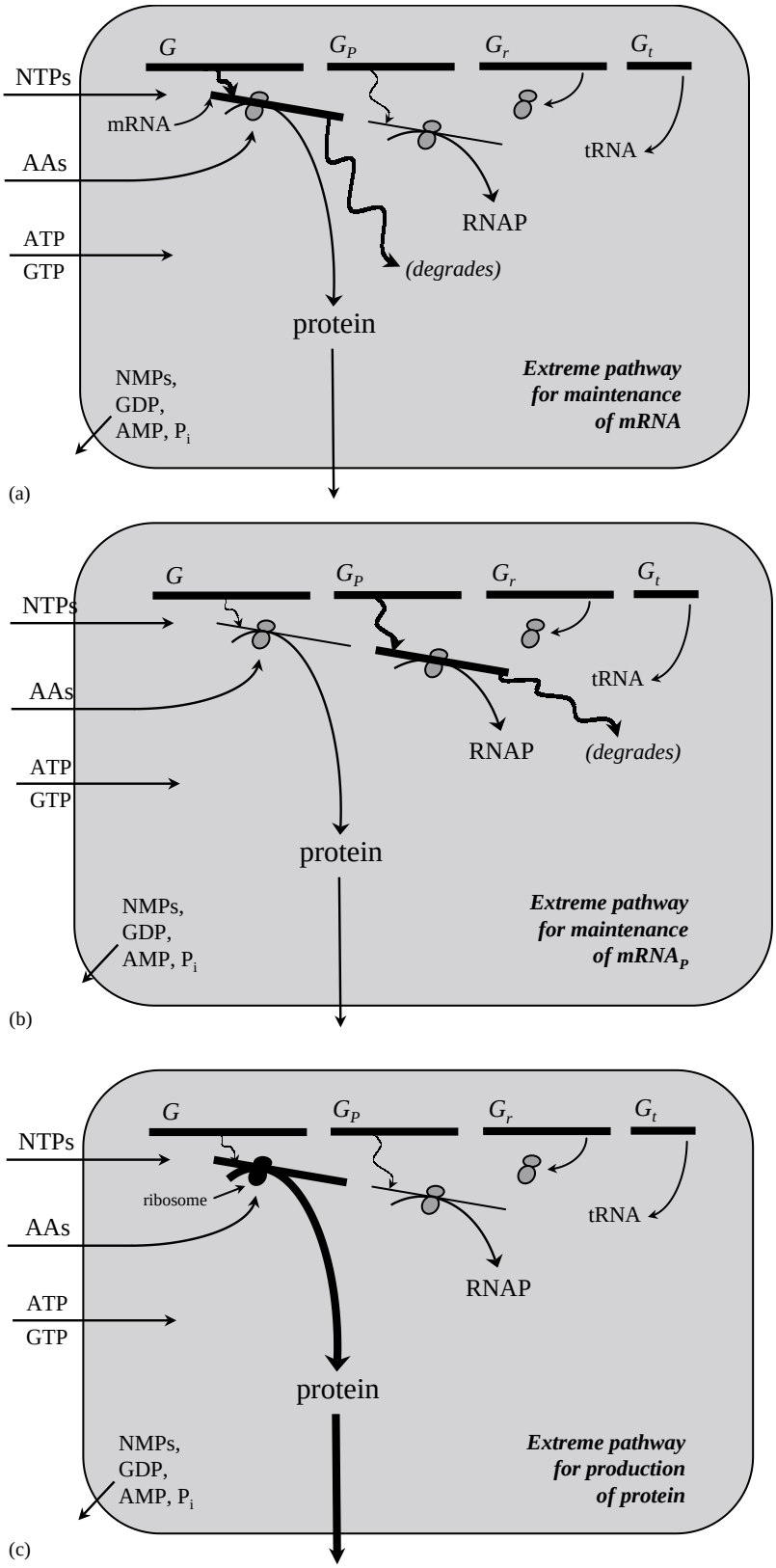

FIG. 4. The extreme pathway structure when the synthesis of internal accessory components are included. There are three extreme pathways in this network: (a) The extreme pathway for the maintenance of mRNA, (b) the extreme pathway for the maintenance of the mRNA encoding for RNAP ($mRNA_P$), and (c) the extreme pathway for the synthesis of protein.

CASE 4—BIOLOGICAL NUMBER OF NTs AND AAs

If we increase the types of nucleotides and amino acids to 4 and 20, respectively, as found in most living cells, the resulting system still contains one maintenance pathway per mRNA and one utilization pathway per protein produced. The extreme pathway analysis thus remains essentially unchanged from Case 3. We still have $N + 1$, albeit more complicated, extreme pathways per N-gene operon (Table 4). Increasing the number of types of nucleotides or amino acids in the system has no effect on the extreme pathway structure, and the limiting amino acid influx to the system determines the protein production flux if the ribosomal pool is not limiting.

CASE 5—PRODUCTION OF *E. COLI* MALATE DEHYDROGENASE

The nucleotide sequence for *mdh* and for the resulting amino acid sequence for malate dehydrogenase are given in Table 5 for *E. coli*. The mRNA encoded by this gene contains 219 adenine, 228 cytosine, 263 guanine, and 229 uracil residues, and the resulting malate dehydrogenase protein consists of the amino acid residues listed in Table 6. The inputs to the simplified system were increased to four types of nucleotide and 20 types of amino acid, and the *E. coli* gene, *mdh*, was selected as G (Blattner *et al.*, 1997). This network exhibits two extreme pathways: a maintenance pathway for $mRNA_{mdh}$ and a protein synthesis pathway for the production of the gene product, malate dehydrogenase. This result is analogous to that in Case 4 for $N = 1$, except that we now have numerical values that correspond to an actual gene in *E. coli*.

Figure 5 provides a schematic of the material costs (i.e. the NTP and AA inputs) and the energy costs (ATP and GTP required) for the maximal production of malate dehydrogenase. If all amino acid influxes are arbitrarily constrained to 10 (units of moles per cell per time) and if the ribosomal saturation is not limiting, then the maximal production flux of malate dehydrogenase is equal to 0.278 mol per cell per time [Fig. 5(a)]. Note that this value is not intended to be a calculation of malate

TABLE 5

The nucleotide sequence of the gene mdh *in* E. coli, *and the translated amino acid sequence of the corresponding malate dehydrogenase protein*

mdh *nucleotide sequence*

```
atgaaagtcgcagtcctcggcgctgctggcggtattggccaggcgcttgcactactgtta
aaaacccaactgccttcaggttcagaactctctctgtatgatatcgctccagtgactccc
ggtgtggctgtcgatctgagccatatccctactgctgtgaaaatcaaaggtttttctggt
gaagatgcgactccggcgctggaaggcgcagatgtcgttcttatctctgcaggcgtagcg
cgtaaaccgggtatggatcgttccgacctgtttaacgttaacgccggcatcgtgaaaaac
ctggtacagcaagttgcgaaaacctgcccgaaagcgtgcattggtattatcactaacccg
gttaacaccacagttgcaattgctgctgaagtgctgaaaaaagccggtgtttatgacaaa
aacaaactgttcggcgttaccacgctggatatcattcgttccaacacctttgttgcggaa
ctgaaaggcaaacagccaggcgaagttgaagtgccggttattggcggtcactctggtgtt
accattctgccgctgctgtcacaggttcctggcgttagttttaccgagcaggaagtggct
gatctgaccaaacgcatccagaacgcgggtactgaagtggttgaagcgaaggccggtggc
gggtctgcaaccctgtctatgggccaggcagctgcacgttttggtctgtctctggttcgt
gcactgcagggcgaacaaggcgttgtcgaatgtgcctacgttgaaggcgacggtcagtac
gcccgtttcttctctcaaccgctgctgctgggtaaaaacggcgtggaagagcgtaaatct
atcggtaccctgagcgcatttgaacagaacgcgctggaaggtatgctggatacgctgaag
aaagatatcgccctgggcgaagagttcgttaataagtaa
```

Malate dehydrogenase amino acid sequence

```
MKVAVLGAAGGIGQALALLLKTQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVKIKGFSG
EDATPALEGADVVLISAGVARKPGMDRSDLFNVNAGIVKNLVQQVAKTCPKACIGIITNP
VNTTVAIAAEVLKKAGVYDKNKLFGVTTLDIIRSNTFVAELKGKQPGEVEVPVIGGHSGV
TILPLLSQVPGVSFTEQEVADLTKRIQNAGTEVVEAKAGGGSATLSMGQAAARFGLSLVR
ALQGEQGVVECAYVEGDGQYARFFSQPLLLGKNGVEERKSIGTLSAFEQNALEGMLDTLK
KDIALGEEFVNK
```

TABLE 6
Amino acid composition of malate dehydrogenase, triosephosphate isomerase, and the proteins encoded by the lac *operon, and the nucleotide composition of the corresponding genes*

	mdh	*lacZ*	*lacY*	*lacA*
A	219	678	240	189
C	228	842	284	125
G	263	888	297	137
T	229	667	433	161
Total	939	3075	1254	612
Ala	35	77	35	8
Arg	8	66	12	10
Asn	11	47	16	16
Asp	12	64	6	9
Cys	3	16	8	2
Gln	14	58	11	0
Glu	20	62	11	14
Gly	36	71	36	16
His	2	34	4	8
Ile	17	39	33	17
Leu	33	96	54	9
Lys	21	20	12	10
Met	4	24	14	7
Phe	10	38	56	8
Pro	13	62	12	12
Ser	17	60	29	12
Thr	18	56	19	12
Trp	0	39	6	2
Tyr	4	31	14	9
Val	34	64	29	22
Total	312	1024	417	203

dehydrogenase actually produced in an *E. coli* cell (i.e. the units are arbitrary), but rather to highlight the limiting constraints on the network that synthesizes this enzyme. Glycine is therefore the limiting amino acid in this scenario since it is the most abundant amino acid in the malate dehydrogenase protein.

Next, we considered the scenario depicted in Fig. 5(b), in which the ribosomal pool limits the overall protein production. In this example, 0.2 is the upper bound on $v_{4\text{-global}}$ (i.e. $\beta_{4\text{-}global\ max} = 0.2$ mol per cell per time). Thus, the maximum attainable protein production flux is also 0.2, and no amino acid is limiting.

CASE 6—THE *LAC* OPERON

Expression of the *lac* operon involves four extreme pathways: one for the maintenance of the polycistronic mRNA, and the other three for the production of each of the three *lac* proteins. This case is analogous to Case 4 with $N = 3$. A schematic of the production of these three proteins is given in Fig. 6(a), and the nucleotide and amino acid composition summaries are provided in Table 6.

If, as before, we constrain all amino acid influxes equally (and if the ribosomal saturation flux is left unconstrained), the limiting factor is the supply of leucine [Fig. 6(b)]. This system can also be constrained by the ribosomal capacity or by any other amino acid if the amino acids are available in uneven supply.

Discussion

We have demonstrated that it is possible to perform a stoichiometry-dependent structural analysis of gene expression and protein synthesis using extreme pathway and flux balance analyses. This framework provides a simple, clear, and detailed accounting of a cell's energy and material expenditure for protein synthesis. Furthermore, the fundamental model presented here is completely scalable, and can readily be expanded to genome-scale via direct use of genomic sequence data.

There are essentially two types of extreme pathways involved in protein production: those involved in the *maintenance* of messenger RNA and those involved in the *utilization* of mRNA to synthesize protein. From a structural standpoint, these processes are decoupled, save that both pathway types are active if the corresponding gene is being expressed and its protein is being synthesized. Thus, their interdependence is strictly logical, whereas the actual flux values are determined by the following key parameters: the mRNA maintenance fluxes (for a particular gene or operon) are strictly dependent upon the promoter strength of the gene under a given set of environmental conditions, except in the event that nucleotide or RNAP availability is limited. The resulting mRNA concentration can then be calculated directly if the half-life of the mRNA is known (Holstege *et al.*, 1998; Cao & Parker, 2001). For the cases studied, the fluxes involved in the utilization of expression information are dependent upon the total ribosomal pool or

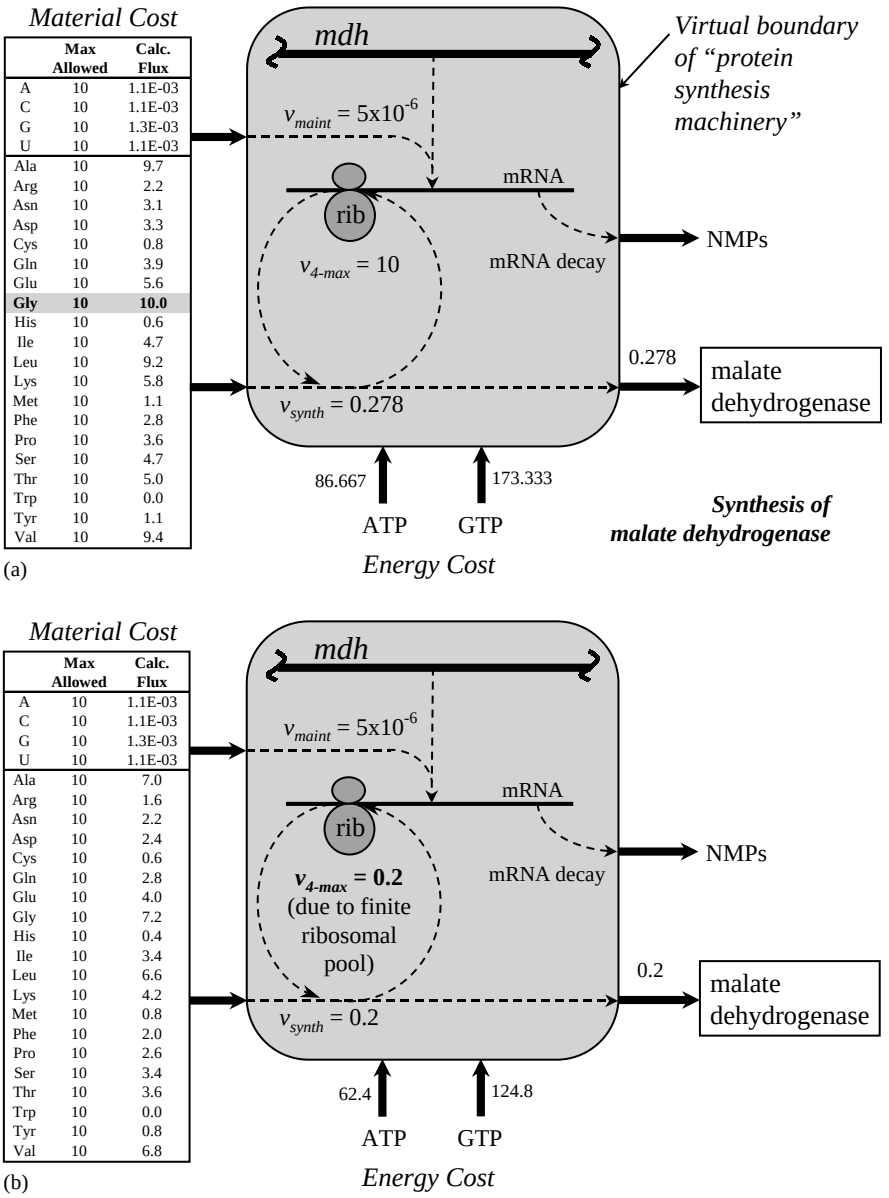

FIG. 5. The synthesis of malate dehydrogenase in *E. coli*. The table on the left of each figure provides the constraints placed upon the nucleotide and amino acid influxes, as well as the calculated influxes upon optimization of protein production. The v_{maint} flux (equal to the $v_{1,2,3}$ flux in the text) was arbitrarily set to 5×10^{-6} (in units of concentration per time). The $v_{4\text{-}max}$ flux, which corresponds to the maximal ribosomal binding flux due to a finite ribosomal pool, is set as a constraint. The v_{synth} is the protein production flux which is being maximized. In panel (a), all amino acid influxes are constrained to 10, and the $v_{4\text{-}max}$ flux is also constrained to 10. In this case, the glycine influx is limiting. In panel (b), the amino acid constraints are unchanged, but the $v_{4\text{-}max}$ flux is constrained to 0.2. The ribosomal pool thus becomes limiting in this case.

upon the availability of amino acids, whichever is limiting.

We have defined the properties of stoichiometric models for individual genes and operons. When we scale-up to describe the protein production of an entire genome, we need to deal with the interactions between these genes (and operons) and the machinery within the fundamental system considered herein. These interactions arise since all genes compete for a finite pool of available RNAP and ribosomes. Thus, the resulting mRNA maintenance fluxes are

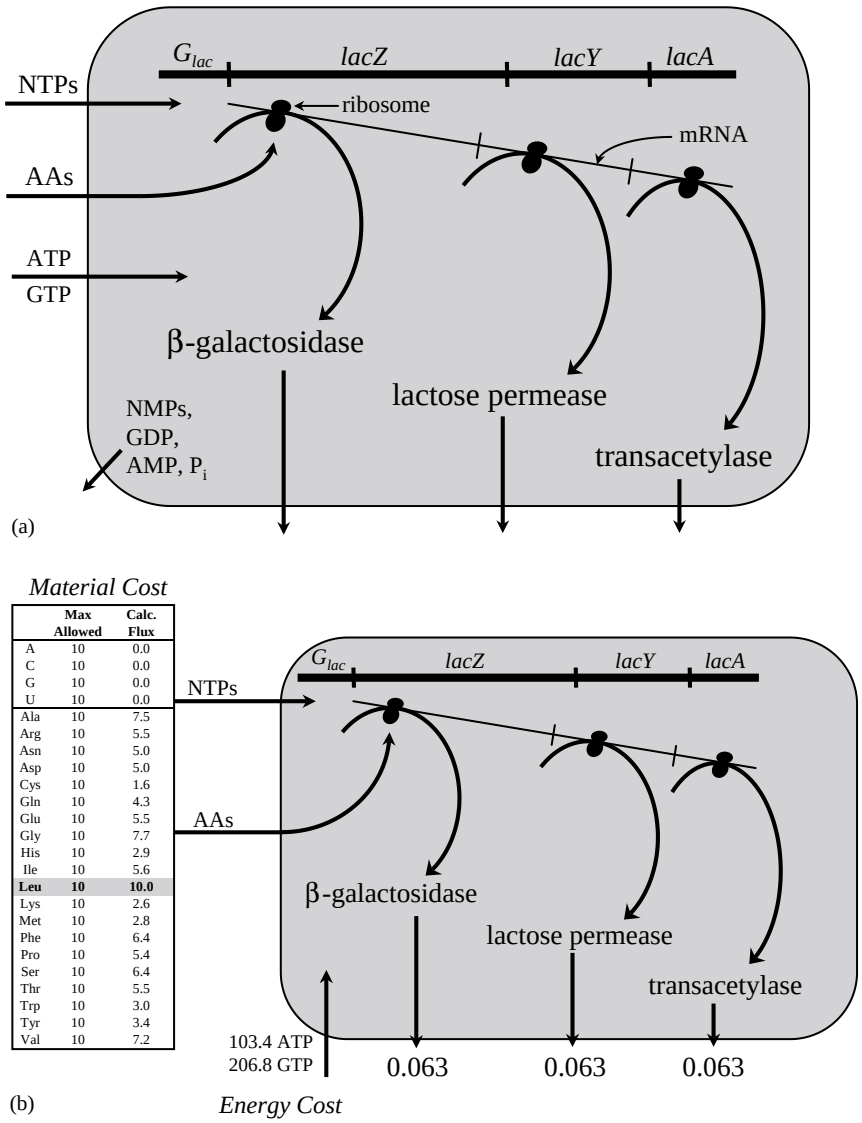

FIG. 6. The expression of the *lac* operon. A simplified schematic for the synthesis of the proteins encoded by the *lac* operon in *E. coli* is provided in panel (a). The mRNA for this operon is polycistronic, encoding for three proteins. In panel (b), the production fluxes of three proteins encoded by the *lac* operon in *E. coli* are being maximized, and the ribosomal binding flux is not limiting. All amino acid influxes are constrained to 10, and the influx of leucine is limiting.

weighted according to the promoter strengths of the various genes. Similarly, the different mRNA transcripts must compete for a limited number of ribosomal binding sites. These translation initiation fluxes must therefore be weighted according to the relative abundances of each mRNA, which, in turn, can be calculated from the corresponding mRNA maintenance fluxes and half-lives [eqn (5)]. In the light of the scarcity of available large-scale promoter strength and mRNA half-life data, the weighting on the translation initiation (i.e. the relative mRNA abundances) fluxes may be estimated directly from gene expression profiles (Tao *et al.*, 1999; Wei *et al.*, 2001).

The overall simplicity of the topology of the reactions involved in protein production is noteworthy in the light of the complexity that has been found to exist in metabolic networks (Schilling & Palsson, 2000; Papin *et al.*, 2002). Here, a lack of robustness is evident in that there are really no choices that can be made within the

mRNA expression maintenance and mRNA expression utilization extreme pathways. The external environment provides a set of inputs that set the fluxes in a condition-dependent manner, and the corresponding proteins are produced from available amino acids and currency metabolites (i.e. ATP and GTP). Thus, the protein synthesis network is more rigid than metabolism.

As genome sequences continue to become available and their gene products are elucidated (the ''parts catalogue'' of the cell, as it were), it is becoming increasingly evident that the interaction of simple components yields tremendous complexity in biology (Palsson, 1997; Strothman, 1997; Alon *et al.*, 1999; Eisenberg *et al.*, 2000). We currently know most of the protein components that are encoded within a genome, and here we have described a fundamental network for the synthesis of each of these proteins. The task at hand is to scale these fundamental networks to include all protein components in a genome, and then to integrate these components with the existing genome-scale metabolic networks, and corresponding regulatory networks when they become available (Covert *et al.*, 2001b).

Taken together, the results presented in this study show that the constraints-based approach of FBA can be used to describe protein synthesis. This approach is readily scaled-up to describe the activity of an entire bacterial genome, and can be integrated with metabolic FBA models.

The authors wish to acknowledge Markus Herrgard for his insightful comments and contributions to this study. Support for this work is provided by grants from the National Science Foundation (BES 98-14092, MCB 98-73384, and BES 01-20363) and the National Institutes of Health (GM-57089).

REFERENCES

Alon, U., Surette, M. G., Barbai, N. & Leibler, S. (1999). Robustness in bacterial chemotaxis. *Nature* **397,** 168–171.

Arkin, A., Ross, J. & McAdams, H. H. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells. *Genetics* **149,** 1633–1648.

Bailey, J. E. (2001). Complex biology with no parameters. *Nat. Biotechnol.* **19,** 503–504.

Blattner, F. R., Plunkett III, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. & Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453–1474.

Bonarius, H. P. J., Schmid, G. & Tramper, J. (1997). Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends Biotechnol.* **15,** 308–314.

Cao, D. & Parker, R. (2001). Computational modeling of eukaryotic mRNA turnover. *RNA* **7,** 1192–1212.

Chvátal, V. (1983). *Linear Programming*. New York: Freeman.

Covert, M. W., Schilling, C. H., Famili, I., Edwards, J. S., Goryanin, I. I., Selkov, E. & Palsson, B. O. (2001a). Metabolic modeling of microbial strains in silico. *Trends Biochem. Sci.* **26,** 179–186.

Covert, M. W., Schilling, C. H. & Palsson, B. O. (2001b). Regulation of gene expression in flux balance models of metabolism. *J. theor. Biol.* **213,** 73–88.

Deuschle, U., Kammerer, W., Gentz, R. & Bujard, H. (1986). Promoters of *Escherichia coli*: a hierarchy of in vivo strength indicates alternate structures. *EMBO J.* **5,** 2987–2994.

Drew, D. A. (2001). A mathematical model for prokaryotic protein synthesis. *Bull. Math. Biol.* **63,** 329–351.

Edwards, J. S. & Palsson, B. O. (1998). How will bioinformatics influence metabolic engineering? *Biotechnol. Bioeng.* **58,** 162–169.

Edwards, J. S. & Palsson, B. O. (1999). Properties of the *Haemophilus influenzae* Rd metabolic genotype. *J. Biol. Chem.* **274,** 17410–17416.

Edwards, J. S. & Palsson, B. O. (2000). The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl Acad. Sci. U.S.A.* **97,** 5528–5533.

Edwards, J. S., Ramakrishna, R., Schilling, C. H. & Palsson, B. O. (1999). Metabolic flux balance analysis. In: *Metabolic Engineering* (Lee, S. Y. & Papoutsakis, E. T., eds). New York: Marcel-Dekker.

Edwards, J. S., Ibarra, R. U. & Palsson, B. O. (2001). In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* **19,** 125–130.

Eisenberg, D., Marcotte, E. M., Xenartos, I. & Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature* **405,** 823–826.

Famili, I., Förster, J., Nielsen, J. & Palsson, B. O. Systems properties of a reconstructed genome-scale metabolic network for *Saccharomyces cerevisiae* (in review).

Fell, D. A. (2001). Beyond genomics. *Trends Genet.* **17,** 680–682.

Förster, J., Famili, I., Fu, P. C., Palsson, B. O. & Nielsen, J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network (in review).

Gombert, A. K. & Nielsen, J. (2000). Mathematical modelling of metabolism. *Curr. Opin. Biotechnol.* **11,** 180–186.

Gotta, S. L., Miller Jr, O. L. & French, S. L. (1991). Ribosomal RNA transcription rate in *Escherichia coli*. *J. Bacteriol.* **173,** 6647–6649.

Holstege, F. C. P., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S. & Young, R. A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95,** 717–728.

Iost, I. & Dreyfus, M. (1995). The stability of *Escherichia coli lacZ* mRNA depends upon the simultaneity of its synthesis and translation. *EMBO J.* **14,** 3252–3261.

Iwakura, Y., Ito, K. & Ishihama, A. (1974). Biosynthesis of RNA polymerase in *Escherichia coli*. I. Control of RNA polymerase content at various growth rates. *Mol. Gen. Genet.* **133,** 1–23.

Kajitani, M. & Ishihama, A. (1983). Determination of the promoter strength in the mixed transcription system. II. Promoters of ribosomal RNA, ribosomal protein S1 and recA protein operons from *Escherichia coli*. *Nucleic Acids Res.* **11,** 3873–3888.

Karp, P. D., Ouzounis, C., & Paley, S. (1996). HinCyc: a knowledge base of the complete genome and metabolic pathways of *H. influenzae*. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **4,** 116–124.

Kushner, S. R. (1996). mRNA decay. In: *Escherichia coli and Salmonella*: *Cellular and Molecular Biology* (Neidhardt, F. C., *et al.*, eds). Washington: ASM Press.

Laffend, L. & Shuler, M. L. (1994). Ribosomal protein limitations in *Escherichia coli* under conditions of high translational activity. *Biotechnol. Bioeng.* **43,** 388–398.

Neidhardt, F. C., Ingraham, J. L. & Schaechter, M. (1990). *Physiology of the Bacterial Cell: A Molecular Approach*. Sunderland, MA: Sinauer.

Nierlich, D. P. (1978). Regulation of bacterial growth, RNA, and protein synthesis. *Annu. Rev. Microbiol.* **32,** 393–432.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27,** 29–34.

Overbeek, R., Larsen, N., Pusch, G. D., D'Souza, M., Selkov, E. Jr., Kyrpides, N., Fonstein, M., Maltsev, N. & Selkov, E. (2000). WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* **28,** 123–125.

Palsson, B. O. (1997). What lies beyond bioinformatics? *Nat. Biotechnol.* **15,** 3–4.

Palsson, B. O. (2000). The challenges of in silico biology. *Nat. Biotechnol.* **18,** 1147–1150.

Papin, J. A., Price, N. D., Edwards, J. S. & Palsson, B. O. (2002). The genome-scale metabolic extreme pathway structure in *Haemophilus influenzae* shows significant network redundancy. *J. theor. Biol.* **215,** 67–82.

Peretti, S. W. & Bailey, J. E. (1986). Mechanistically detailed model of cellular metabolism for glucose-limited growth of *Escherichia coli* B/r-A. *Biotechnol. Bioeng.* **28,** 1672–1689.

Record Jr, M. T. Reznikoff, W. S., Craig, M. L., McQuade, K. L. & Schlax, P. J. (1996). *Escherichia coli* RNA polymerase ($E\sigma^{70}$), promoters, and the kinetics of the steps of transcription initiation. In: *Escherichia coli and Salmonella: Cellular and Molecular Biology* (Neidhardt, F. C., *et al.*, eds). Washington: ASM Press.

Santillán, M. & Mackey, M. C. (2001). Dynamic regulation of the tryptophan operon: a modeling study and comparison with experimental data. *Proc. Natl Acad. Sci. U.S.A.* **98,** 1364–1369.

Schilling, C. H. & Palsson, B. O. (2000). Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. *J. theor. Biol.* **203,** 249–283.

Schilling, C. H., Edwards, J. S. & Palsson, B. O. (1999). Toward metabolic phenomics: analysis of genomic data using flux balances. *Biotechnol. Prog.* **15,** 288–295.

Schilling, C. H., Letscher, D. & Palsson, B. O. (2000). Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. theor. Biol.* **203,** 229–248.

Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S. & Palsson, B. O. Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.* **184.**

Selkov Jr, E., Grechkin, Y., Mikhailova, N. & Selkov, E. (1998). MPW: the metabolic pathways database. *Nucleic Acids Res.* **26,** 43–45.

Sinha, S. (1988). Theoretical study of the tryptophan operon: application in microbial technology. *Biotechnol. Bioeng.* **31,** 117–124.

Strothman, R. C. (1997). The coming Kuhnian Revolution in biology. *Nat. Biotechnol.* **15,** 194–199.

Tao, H., Bausch, C., Richmond, C., Blattner, F. R. & Conway, T. (1999). Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. *J. Bacteriol.* **181,** 6425–6440.

Thomas, R. (1991). Regulatory networks seen as asynchronous automata: a logical description. *J. theor. Biol.* **153,** 1–23.

Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T. S., Matsuzaki, Y., Miyoshi, F., Saito, K., Tamida, S., Yugi, K., Venter, J. C. & Hutchison, C. A., III (1999). E-CELL: software environment for whole-cell simulation. *Bioinformatics* **15,** 72–84.

Varma, A. & Palsson, B. O. (1993). Metabolic capabilities of *Escherichia coli*: II. Optimal growth patterns. *J. theor. Biol.* **165,** 503–522.

Varma, A. & Palsson, B. O. (1994). Metabolic flux balancing: basic concepts, scientific and practical use. *Bio/Technology* **12,** 994–998.

Vogel, U. & Jensen, K. F. (1994). The RNA chain elongation rate in *Escherichia coli* depends on the growth rate. *J. Bacteriol.* **176,** 2807–2813.

Wagner, R. (2000). *Transcription Regulation in Prokaryotes*. New York: Oxford.

Wei, Y., Lee, J.-M., Richmond, C., Blattner, F. R., Rafalski, J. A. & La Rossa, R. A. (2001). High-density microarray-mediated gene expression profiling of *Escherichia coli*. *J. Bacteriol.* **183,** 545–556.

Weller, K. & Recknagel, R. D. (1994). Promoter strength prediction based on occurrence frequencies of consensus patterns. *J. theor. Biol.* **171,** 355–359.

Wong, P., Gladney, S., & Keasling, J. D. (1997). Mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. *Biotechnol. Prog.* **13,** 132–143.

Zacharias, M., Theissen, G., Bradaczek, C. & Wagner, R. (1991). Analysis of sequence elements important for the synthesis and control of ribosomal RNA in *E. coli*. *Biochimie* **73,** 699–712.

Appendix A

Abbreviations and symbols used

a	number of amino acids in the protein of the simplified one-amino acid system
AA	generic amino acid
AA-tRNA	generic aminoacyl transfer RNA
G	an arbitrary gene
G_P	gene encoding for RNAP
G_r	gene encoding for ribosomal RNA
G_t	gene encoding for transfer RNA
G^*	gene with bound RNAP (open-promoter complex)
mRNA	messenger RNA
N	number of genes in an operon in the sample cases studied
n	number of nucleotides in the mRNA of the simplified one-nucleotide system
NMP	generic nucleotide monophosphate
NTP	generic nucleotide triphosphate
$\mathbf{P}_i$	inorganic phosphate
$\mathbf{PP}_i$	pyrophosphate
protein	generic protein
rib	ribosome
rib*	ribosome–mRNA complex
RNAP	RNA polymerase
S	stoichiometric matrix
tRNA	transfer RNA
$\mathbf{v}$	flux vector