A scheme for the analysis of microarray measurements based on a quantitative theoretical framework for bacterial cell growth: application to studies of Mycobacterium tuberculosis

Robert A. Cox
Division of Mycobacterial Research, National Institute for Medical Research, London NW7 1AA, UK

A theoretical framework was established for the interpretation of microarray measurements. Mathematical equations were derived that link the molecular processes involved in the transcription and translation of an open reading frame (ORF) with the properties of a population of cells. The theory was applied to three published sets of microarray measurements related to the growth of Mycobacterium tuberculosis. It was shown for strains growing at the same rate, for example wild-type and mutant strains, that the expression ratio obtained by microarray analysis for a particular ORF is equal to the ratio of the copy numbers of the encoded protein. The growth of M. tuberculosis in a batch culture was analysed at several time points over a period of 60 days. Several properties including the following were calculated for cells cultured for 60 days: \( \mu \leq 0.008 \text{ h}^{-1} \), there was a decrease in the number of ribosomes per cell to 26% of the value at day 0, and only 40% or less of this reduced number of ribosomes were estimated to be actively synthesizing protein. Profiles of the expression ratio observed for a particular ORF versus the period of cell culture were related to changes in the relative numbers of copies of the encoded protein per cell. Two profiles were found to have theoretical significance: profile I, exemplified by ORFs encoding proteins needed for DNA partition and DNA synthesis; and profile II, exemplified by ORFs encoding proteins (including ribosomal proteins) needed for protein synthesis. Data for a number of other genes including hspX, icl, dosR and ftsZ were also analysed.

INTRODUCTION

Current methods of studying gene expression include the use of reverse transcriptase to make cDNA copies of cellular RNA and assays for particular gene transcripts based on either DNA amplification (quantitative real-time polymerase chain reaction: qRT-PCR) or on hybridization of tagged cDNA to an array of immobilized DNA targets, with each target representing an individual gene (microarray analysis). The results of qRT-PCR are expressed as the number of copies of transcript per unit of RNA; frequently a reference gene is also included as a ‘normalizer’ to place the expression of the gene of interest in a cellular context. Microarray analysis is generally used as a comparative tool to ascertain whether a change in either growth conditions or genotype leads to ‘upregulation’ or ‘downregulation’ of sets of genes of interest. Both techniques have proved to be valuable and widely used methods for studying the expressions of individual genes. Mycobacterial gene expression has been studied by both qRT-PCR and microarray analysis. Thus far, the expression ratios (the output from microarray measurements) have not been directly related to cell parameters. Previously (Cox, 2004), I presented a theoretical framework for bacterial cell growth based on the properties of population-average cells and the mathematical relationships involved in transcription and translation. In the bacterial cell the latter two properties are coupled (Stent, 1964; Byrne et al., 1964; Miller et al., 1970). I now show that this framework can be readily applied to the interpretation of microarray data. This paper describes how parameters of population-average cells are related to microarray measurements; and how microarray data may be used to reveal relative values of properties such as the number of copies of a particular protein per cell and the average number of ribosomes per cell.

The procedures developed were applied to three microarray studies (Bacon et al., 2004; Kendall et al., 2004; Voskuil et al., 2004; for review see Butcher, 2004) of the growth of Mycobacterium tuberculosis. Features of the synthesis of proteins, DNA and RNA were identified and changes in the physiological parameters of the tubercle bacillus during growth were quantified.

Abbreviations: qRT-PCR, quantitative real-time polymerase chain reaction; RNAP, RNA polymerase.
THEORY

Definition of the problem. The problem is to find a theoretical framework that links experimental data with cell parameters that are accessible to experiment. Microarray analysis is a comparative technique. Usually a reference strain (identified by a single superscript prime) and an experimental strain (identified by two superscript primes) are compared. RNA is isolated from each strain. A sample, x μg, of each RNA fraction is copied into cDNA using dNTPs tagged with fluorophors f and f’ respectively. Samples of cDNA’ and cDNA” are mixed, denatured and competitively hybridized to an array of DNA sequences immobilized on a glass surface. An all-genome array is constructed so that each open reading frame (ORF) is represented by a specific DNA sequence at a known site. The fluorescence measurements are normalized, allowing calculation of the expression ratio \( r_{(i)} = \frac{f''_{(i)}}{f'_{(i)}} \) of the two fluorescent labels hybridized to a particular ORF \( \text{ORF}_{(i)} \). Provided that cDNA’ and cDNA” both accurately reflect the composition of their parent RNA fraction, \( r_{(i)} \) then provides a measure of the expression of \( \text{ORF}_{(i)} \) under reference and experimental conditions (Bowtell & Sambrook, 2000).

The coupling of transcription and translation that is a characteristic feature of bacterial growth provides the basis for the theoretical framework that is sought. Thus, the problem can be restated as a search for the appropriate cell parameters that define the expression ratio \( r_{(i)} \). As a first step it is helpful to show how the properties of a cell are related to the properties of a culture. See Table 1 for definitions of variables.

Table 1. Definitions of variables

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition of variable (units in parentheses)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Cell age [fraction of the generation time (time from birth/generation time); ( a=0 ) for a newborn cell and ( a=1 ) for a cell about to divide]</td>
</tr>
<tr>
<td>( D_{(i)} )</td>
<td>Mass (fg) of DNA per ml culture at time ( t ) (h)</td>
</tr>
<tr>
<td>f</td>
<td>Fluorescence of cDNA hybridized to immobilized DNA</td>
</tr>
<tr>
<td>( l_{at(i)} )</td>
<td>Length (aa) of a protein ( p_{(i)} ) encoded by ( \text{ORF}_{(i)} )</td>
</tr>
<tr>
<td>( M_{(i)} )</td>
<td>Mass of transcripts (fg mRNA(<em>{(i)})) of ( \text{ORF}</em>{(i)} ) per x μg RNA (see Fig. 1)</td>
</tr>
<tr>
<td>( m'_{\text{nuc}} )</td>
<td>Average molecular mass of a nucleotide</td>
</tr>
<tr>
<td>( m_p_{(av)\dagger} )</td>
<td>Mass of protein per population-average cell (fg)</td>
</tr>
<tr>
<td>( m_{RNA_{(av)}}\dagger )</td>
<td>Mass of RNA per population-average cell (fg)</td>
</tr>
<tr>
<td>( m_{mRNA_{(av)}}\dagger )</td>
<td>Mass of mRNA per population-average cell (fg)</td>
</tr>
<tr>
<td>( m_{rRNA_{(av)}}\dagger )</td>
<td>Mass of rRNA per population-average cell (fg)</td>
</tr>
<tr>
<td>( m_{tRNA_{(av)}}\dagger )</td>
<td>Mass of tRNA per population-average cell (fg)</td>
</tr>
<tr>
<td>( n_{\text{cells}} )</td>
<td>Number of cells per culture</td>
</tr>
<tr>
<td>( n_{c-p(i)\dagger} )</td>
<td>Number of copies of protein ( p_{(i)} ) per population-average cell</td>
</tr>
<tr>
<td>( n_{\text{ORF}/\text{fg}} )</td>
<td>Number of ORFs per genome</td>
</tr>
<tr>
<td>( n_{R_{(av)}}\dagger )</td>
<td>Number of ribosomes per population-average cell</td>
</tr>
<tr>
<td>( n_{R(i)\dagger} )</td>
<td>Number of ribosomes actively translating transcripts of ( \text{ORF}_{(i)} )</td>
</tr>
<tr>
<td>( n_{g(i)\dagger} )</td>
<td>Number of transcript equivalents being synthesized by ( \text{ORF}_{(i)} )</td>
</tr>
<tr>
<td>( p_{(i)} )</td>
<td>Mass of protein (fg) per ml culture at time ( t )</td>
</tr>
<tr>
<td>( R_{(i)} )</td>
<td>Mass of RNA (fg) per ml culture at time ( t )</td>
</tr>
<tr>
<td>( r^{\dagger} )</td>
<td>Expression ratio ( (f''/f') )</td>
</tr>
<tr>
<td>( r_{(i)\dagger} )</td>
<td>Expression ratio for ( \text{ORF}_{(i)} )</td>
</tr>
<tr>
<td>( r_{(i)_{(av)}\dagger} )</td>
<td>Average of the number specified of expression ratios of ORFs encoding proteins with related functions</td>
</tr>
<tr>
<td>( r_{(i)_{(i)\dagger}} )</td>
<td>Expression ratio of ( \text{ORF}_{(i)} ) encoding a protein whose copy number is independent of growth rate</td>
</tr>
<tr>
<td>( r_{(k)\dagger} )</td>
<td>Expression ratio of ( \text{ORF}_{(k)} ) encoding a ribosomal protein</td>
</tr>
<tr>
<td>t</td>
<td>Time (h)</td>
</tr>
<tr>
<td>( t_D )</td>
<td>Generation time (h)</td>
</tr>
<tr>
<td>x</td>
<td>Mass (μg) of RNA used to prepare cDNA</td>
</tr>
<tr>
<td>( \beta_{R(\text{av})\dagger} )</td>
<td>Fraction of ribosomes actively synthesizing protein</td>
</tr>
<tr>
<td>( \varepsilon_{(\text{av})\dagger} )</td>
<td>Average peptide chain elongation rate (amino acids incorporated h(^{-1})) of proteins</td>
</tr>
<tr>
<td>( \varepsilon_{\text{aa}(i)\dagger} )</td>
<td>The peptide chain elongation rate (amino acids incorporated h(^{-1})) of protein ( p_{(i)} )</td>
</tr>
<tr>
<td>( \mu )</td>
<td>Specific growth rate (h(^{-1}))</td>
</tr>
<tr>
<td>( v )</td>
<td>Number of terms in a summation; see equation 11d</td>
</tr>
<tr>
<td>( \Sigma )</td>
<td>Sign indicating summation from ( i=1 ) to ( i=n_{\text{ORF}/\text{fg}} )</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>Average number of mRNA nucleotides per programmed ribosome</td>
</tr>
<tr>
<td>( \psi )</td>
<td>Size (bp) needed to form an open complex with RNAP</td>
</tr>
<tr>
<td>( \omega_{p_{(av)}}\dagger )</td>
<td>Specific protein synthesis rate (fg protein synthesized h(^{-1})) (see equation 2A)</td>
</tr>
<tr>
<td>( \omega_{p_{(i)\dagger}} )</td>
<td>Specific protein synthesis rate (fg of protein ( p_{(i)} ) synthesized h(^{-1})) of protein ( p_{(i)} ) (see equation 3A)</td>
</tr>
</tbody>
</table>

\( \dagger \) Symbols marked with a dagger may be modified either by a superscript prime (e.g. \( n'_{c-p(i)} \)) or by a superscript double prime (e.g. \( n''_{c-p(i)} \)). The superscript prime denotes reference cells and the superscript double prime denotes 'experimental' cells; that is either mutant cells or a change in the conditions of cell culture. The dagger signifies a parameter important to equation 11 that is a function of \( \mu \).
Population-average cells. Suppose that at t hours a bacterial culture growing exponentially comprises n_{cell}(t) cells and the masses of protein, RNA, DNA, etc., are p(t), R(t) and D(t), etc. A population-average cell (Schaechter et al., 1958) has the properties m_{p(avg)} fg protein, m_{RNA(avg)} fg RNA, m_{DNA(avg)} fg DNA and so on as defined in equations 1, 2 and 3, where the subscript (av) specifies a population-average cell.

\[ P(t) = n_{cell}(t) \cdot m_{p(avg)} \]  
(1)

\[ R(t) = n_{cell}(t) \cdot m_{RNA(avg)} \]  
(2)

\[ D(t) = n_{cell}(t) \cdot m_{DNA(avg)} \]  
(3)

The population-average cell reflects the properties of the entire cell population, which includes cells of all ages, a, ranging from newborn cells (a=0) to cells about to divide (a=1); see Cox (2003, 2004) and Cox & Cook (2007). In an individual cell particular proteins may be expressed at specific stages of cell growth. All such proteins are present within population-average cells although the element of periodic expression is not recognized. By definition, the abundances of such proteins within the cell culture will increase at the same rate as population-average cells. The quantity measured in microarray analysis may now be defined in terms of parameters of population-average cells.

What is the quantity measured in microarray analysis? Let prime and double prime superscripts denote the two fluorescently labelled cDNA preparations and also the cell parameters related to the RNA preparations from which the cDNA samples were prepared. For example, \( M_{tr(i)}' \) and \( M_{tr(i)}'' \) each represent the mass of transcripts of ORF(j) per \( x \) \( \mu \)g RNA used as the substrate for cDNA synthesis. Thus \( r_0(t) \) the relative intensities of the two fluorescent labels, is given by equation 4:

\[ r_0(t) = M_{tr(i)'} / M_{tr(i)''} \]  
(4)

Equation 4, which is fundamental to microarray analysis, is based on three assumptions: first, the fluorescence arising from hybridization to immobilized DNA_{av} is a measure of the concentration of cDNA_{av} in solution; secondly, the ratio \( r_0(t) \) of the fluorescent labels is equal to the ratio \( M_{tr(i)'} / M_{tr(i)''} \); thirdly, the ratio \( r_0(t) \) is unchanged by hybridization. However, \( M_{tr(i)'} \) and \( M_{tr(i)''} \) are related to the cell parameters \( m'_{RNA(avg)} \) and \( m''_{RNA(avg)} \), which are the masses (fg) of mRNA(i) per cell, and \( m'_{RNA(avg)} \) and \( m''_{RNA(avg)} \) as shown in equations 5 and 6:

\[ M_{tr(i)'} = m'_{RNA(avg)}(x/m'_{RNA(avg)}) \]  
(5)

\[ M_{tr(i)''} = m''_{RNA(avg)}(x/m''_{RNA(avg)}) \]  
(6)

The terms \( x/m'_{RNA(avg)} \) and \( x/m''_{RNA(avg)} \) reveal the numbers of cells needed to yield \( x \) \( \mu \)g RNA. Substitution for \( M_{tr(i)'} \) and \( M_{tr(i)''} \) in equation 4 leads to equation 7:

\[ r_0(t) = m'_{RNA(avg)} / m''_{RNA(avg)} \]  
(7)

The next step is to identify the cell parameters governing \( m_{RNA(avg)} \).

Transcription/translation of an individual ORF (ORF(i)). A snapshot of an ORF undergoing transcription/translation has revealed (Miller et al., 1970) RNA polymerase molecules (RNAPs) at intervals along the gene, with nascent mRNA transcripts increasing in size according to the position of the RNAP; and a proportionate number of ribosomes attached to each nascent mRNA. The specific synthesis rate \( \omega_p(i) \) of protein \( p(i) \) is the product of the number \( n_{R(i)} \) of ribosomes actively translating mRNA(i) and \( e_{aa(i)} \) the polypeptide chain elongation rate and \( m_e \) the average mass of an amino acid (see equation 8):

\[ \omega_p(i) = m_e \cdot n_{R(i)} \cdot e_{aa(i)} \]  
(8)

It is inferred that mRNA \( R(i) \) is in the process of being translated is protected from RNase action. Suppose that \( \sigma \) nucleotides of mRNA \( R(i) \) are protected by each ribosome, then \( m_{RNA(i)} \) is given by equation 9, where \( m_{nuc} \) is the average mass of a nucleotide:

\[ m_{RNA(i)} = m_{nuc} \cdot \sigma \cdot n_{R(i)} \]  
(9)

It is shown in the Appendix that manipulations of equations 9 and 7A and related parameters lead to equation 10, where \( l_{R(i)} \) is the length in amino acids of protein \( p(i) \) and \( n_{c-p(i)} \) is the number of copies of protein \( p(i) \) per cell:

\[ m_{RNA(i)} = m_{nuc} \cdot \sigma \cdot n_{c-p(i)} \cdot l_{R(i)} / l_{a(i)} \]  
(10)

Substitutions for \( m'_{RNA(avg)} \) and \( m''_{RNA(avg)} \) lead to equation 11a:

\[ r_0(t) = (n'_{c-p(i)} / n''_{c-p(i)}) \cdot (l''_{a(i)} / l'_{a(i)}) \cdot (m'_{RNA(avg)} / m''_{RNA(avg)}) \]  
(11a)

Simplified forms of equation 11a used in this study. Equation 11a relates the expression ratio \( r_0(t) \) to parameters of population-average cells; that is, \( r_0(t) \) is a function of relative values of \( \mu \), \( e_{aa(i)} \), \( m'_{RNA(avg)} \) and \( n_{c-p(i)} \). The solution of equation 11a requires measurements of \( \mu \) and the ratios DNA:protein for each culture. These measurements provide (Cox, 2004) values of \( m'_{RNA(avg)} \), \( m''_{RNA(avg)} \), \( m'_{p(avg)} \) and \( m''_{p(avg)} \). \( e_{aa(i)} \) (note that \( e_{aa(i)} \approx e''_{aa(i)} \) and \( e'_aa(i) \approx e''_{aa(i)} \). There are several situations that allow simplification of equation 11a as discussed below.

Wild-type and mutant strains grow at the same rate. It has been shown (Schaechter et al., 1958) that at a given temperature the macromolecular composition of a bacterium is dependent on growth rate; notably, the RNA (ribosome) content has been found to vary with growth rate. For a review see Bremer & Dennis (1996). Conversely, it can be assumed that a mutation that has no effect on the specific growth rate has no effect on the cell’s macromolecular composition. Suppose that a single prime and double prime respectively denote wild-type and mutant; when \( \mu'' = \mu' \), \( m'_{RNA(avg)} = m''_{RNA(avg)} \), \( e'_{aa(i)} = e''_{aa(i)} \) and hence \( e'_{aa(i)} \approx e''_{aa(i)} \). The expression ratio \( r = r' / r'' \) is then equal to \( n'_{c-p(i)} / n''_{c-p(i)} \). In other words, the expression ratio is equal to the ratio of the copy numbers of protein \( p(i) \) in mutant and wild-type (see equation 11b: Table 2).

The number \( n_c-p(k) \) of copies of protein \( p(k) \) per cell is independent of growth rate. Suppose that the bacterial cell uses a fixed complement \( n_{c-p(k)} \) of protein \( p(k) \) and that the number \( n_{c-p(k)} \) of copies per cell of protein \( p(k) \) encoded by ORF(k) is the same at all growth rates. In this case \( n'_{c-p(k)} / n''_{c-p(k)} = 1 \) and that equation 11a reduces to equation 11c (Table 2). Candidates for proteins such as \( p(k) \) include those involved in processes that take place only during cell division; for example, chromosome segregation. Note that equation 11c has a further property: the ratio \( r_0(t)/r_0 \) is equal to \( n'_{c-p(k)} / n''_{c-p(k)} \).

The number \( n_{c-p(k)} \) of copies of protein \( p(k) \) is directly proportional to the number \( n_{R(avg)} \) of ribosomes per cell. Suppose that ORF(k) encodes protein \( p(k) \) which has a number \( n_{c-p(k)} \) of copies per population-average cell that is directly proportional to \( n_{R(avg)} \) that is, \( n_{c-p(k)} \) is equal to the product of a constant \( K(>0) \) and \( n_{R(avg)} \). In this case, \( n'_{c-p(k)}/n''_{c-p(k)} \) is equal to \( n'_{R(avg)} / n''_{R(avg)} \) which, in turn, approximates to \( m'_{RNA(avg)}/m''_{RNA(avg)} \). Thus, equation 11a is reduced to equation 11c (Table 2). Obvious examples of the \( p(k) \) family of proteins are ribosomal (r-proteins) themselves. L7/L12 is present as four copies per ribosome and all others are present as a single copy per ribosome (Bremer & Dennis, 1996). Except for L7/L12 the copy number \( n_{c-p(k)} \) of an r-protein is equal to \( n_{R(avg)} \). Other components of the translational machinery such as initiation factors, elongation factors and aminoacyl-RNA synthases (Bremer & Dennis, 1996) were found.
to have copy numbers directly proportional to \( n_{R_{(av)}} \). The ratio \( r_{(i)} / r_{(j)} \) yields \( m'_{RNA_{(av)}} / m'_{RNA_{(av)}} \). The equations shown in Table 2 provide the basis for a systematic analysis of the expression of each ORF of a bacterial genome. The equations derived and the assumptions upon which they are based are supported by the microarray studies of \( M. tuberculosis \) gene expression (see below).

### RESULTS AND DISCUSSION

Support for equation 11a and the related forms was sought from published microarray studies of \( M. tuberculosis \). Equation 11a was not tested directly because the additional data needed (\( \mu \) and the ratios DNA : RNA : protein) are not available. However, equation 11b, a simpler form of equation 11a, was applied to two studies; in each study reference and experimental strains were found to grow at the same rate. It has been shown (Verma et al., 1999) that, like other bacteria (Schaechter et al., 1958; Bremer & Dennis, 1996), the tubercle bacillus maintains the number of ribosomes per cell commensurate with the needs of the cell for protein synthesis (growth rate control of ribosome synthesis). It follows that if two strains of \( M. tuberculosis \) grow at the same rate, both will synthesize proteins at the same rate and have similar numbers of ribosomes per cell. Equation 11a then reduces to equation 11b so that the expression ratio \( r_{(i)} \) is equal to \( n'_{c_{(p(i)}} / n'_{c_{(p(j)}} \). In the particular case of ribosomal proteins \( r=1 \) when \( n'_{R_{(av)}} / n'_{R_{(av)}}=1 \) because \( n'_{c_{(p(k)}} / n'_{c_{(p(j)}} = n'_{R_{(av)}} / n'_{R_{(av)}}=1 \). These considerations were applied to two studies as described below in section A.

The results of a third study were examined in order to establish the extent to which equations 11c–11b can be applied to studies where reference strains do not grow at the same rate. Voskuil et al. (2004) cultured \( M. tuberculosis \) over a period of 60 days and used microarray analysis to follow changes in gene expression as nutrients were consumed, leading to a decrease in the specific growth rate (see section B below).

### A. Reference and experimental strains grown at the same rate

Two sets of microarray data were examined. In the first study, it was found that when nitrogen was limiting, the tubercle bacillus grew in a chemostat at the same rate (\( \mu=0.029 \text{ h}^{-1} \)) at two different concentrations of dissolved oxygen (Bacon et al., 2004). In the second study, it was shown that wild-type \( M. tuberculosis \) and a dosR 'knockout' mutant grew at the same rate in batch culture. The results of a third study (Voskuil et al., 2004) provide a control. \( M. tuberculosis \) was grown in batch culture; RNA was isolated from mid-exponential-phase cells and was used as the substrate for the preparation of both the prime and double prime fluorescent cDNA samples. The expression ratios obtained by microarray analysis provide a positive test for equation 11b because only a single strain grown at a particular specific growth rate is involved.

A panel of 50 genes encoding ribosomal proteins (see Table 3) was identified. Six of eight putative ribosomal genes (see Table 4) were excluded because the functions of the encoded proteins are uncertain. There are two candidate genes present in the genome for each of the ribosomal proteins RpsN, RpsR, RpmB and RpmG and

### Table 2. Useful forms of equation 11

<table>
<thead>
<tr>
<th>Eqn no.</th>
<th>Equation</th>
<th>Special conditions</th>
<th>Information yielded</th>
</tr>
</thead>
<tbody>
<tr>
<td>11a</td>
<td>( r_{(i)} = \frac{n'<em>{c</em>{(p(i)}}/n'<em>{c</em>{(p(j)}}}{(m'<em>{RNA</em>{(av)}}/m'<em>{RNA</em>{(av)}})} \cdot (\epsilon'_{aa(i)})/\mu' )</td>
<td>None</td>
<td>( n'<em>{c</em>{(p(i)}}/n'<em>{c</em>{(p(j)}} ) when ( \mu' ), ( \epsilon'<em>{aa(i)} ), ( \epsilon'</em>{aa(j)} ), ( m'<em>{RNA</em>{(av)}} ) and ( m'<em>{RNA</em>{(av)}} ) are known</td>
</tr>
<tr>
<td>11b*</td>
<td>( r_{(i)} = \frac{\mu'}{\mu'} \cdot \frac{n'<em>{c</em>{(p(i)}}}{n'<em>{c</em>{(p(j)}}} \cdot \frac{m'<em>{RNA</em>{(av)}}}{m'<em>{RNA</em>{(av)}}} \cdot \epsilon'_{aa(i)} )</td>
<td>( \mu' = \mu' ), ( \epsilon'<em>{aa(i)} = \epsilon'</em>{aa(j)} )</td>
<td>( n'<em>{c</em>{(p(i)}}/n'<em>{c</em>{(p(j)}} )</td>
</tr>
<tr>
<td>11c†</td>
<td>( r_{(i)} = \frac{\mu'}{\mu'} \cdot \frac{n'<em>{c</em>{(p(i)}}}{n'<em>{c</em>{(p(j)}}} \cdot \frac{m'<em>{RNA</em>{(av)}}}{m'<em>{RNA</em>{(av)}}} \cdot \epsilon'_{aa(i)} )</td>
<td>( \mu' = \mu' ), ( \epsilon'<em>{aa(i)} = \epsilon'</em>{aa(j)} )</td>
<td>( n'<em>{c</em>{(p(i)}}/n'<em>{c</em>{(p(j)}} )</td>
</tr>
<tr>
<td>11d‡</td>
<td>( m'<em>{p</em>{(av)}} / m'<em>{p</em>{(av)}} = \sum_{i} \theta_{i} \frac{\epsilon'<em>{aa(i)}}{\epsilon'</em>{aa(j)}} )</td>
<td>See Appendix</td>
<td>See equation 11f</td>
</tr>
<tr>
<td>11e§</td>
<td>( \mu' \leq \mu'_{(k)} \cdot \mu' )</td>
<td>( (\beta'<em>{m}/\beta'</em>{m}) )</td>
<td>( n'<em>{R</em>{(av)}} / n'<em>{R</em>{(av)}} )</td>
</tr>
<tr>
<td>11f†</td>
<td>( \mu' \leq \mu'_{(k)} \cdot \mu' )</td>
<td>( (\beta'<em>{m}/\beta'</em>{m}) )</td>
<td>( n'<em>{R</em>{(av)}} / n'<em>{R</em>{(av)}} )</td>
</tr>
</tbody>
</table>

\( * \)Equation 11b applies when, for example, cultures of wild-type and a mutant strain are compared and both strains grow at the same rate (see Kendall et al., 2004); or when the same strain is cultured under different growth conditions and grows at the same rate (Bacon et al., 2004).

\( † \)Equation 11c applies when the protein \( p_{(i)} \) encoded by ORF \( _{(i)} \) has the same copy number per cell (\( n_{c_{(p(i)}} \)) under all conditions of growth.

\( ‡ \)Equation 11d is an approximation for \( m'_{p_{(av)}} / m'_{p_{(av)}} = m'_{a_{(av)}} \cdot \sum_{i} \theta_{i} \frac{(\epsilon'_{aa(i)})/(\epsilon'_{aa(j)})}{m'_{RNA_{(av)}}/m'_{RNA_{(av)}}} \) provided that RNA accounts for the same fraction of \( m'_{RNA_{(av)}} \) in each case.

\( †† \)Equation 11f, which is an approximate form of equation 11e, is based on the assumption that \( \epsilon'_{aa(i)} \approx \epsilon'_{aa(j)} \) and \( \epsilon'_{aa(k)} \approx \epsilon'_{aa(l)} \).

\( § \)Equation 11h follows from equation 11e on the basis of the assumption that \( \epsilon'_{aa(i)} / \epsilon'_{aa(l)} \leq \epsilon'_{aa(l)} \).
there is no information to show which of the encoded proteins are incorporated into ribosomes.

The \( r \)-values for the three sets of 50 ribosomal proteins were compared (see Fig. 1); the mean \( r \)-values, \( r_{50} \), and standard deviations were calculated and individual \( r \)-values were found to be in accord with the corresponding normal distribution. Thus, \( r \)-values of each set of 50 ribosomal proteins are accurately represented by \( r_{50} \pm SD \) (standard deviation). Thus, equation 12 summarizes the data for \( M. \) tuberculosis ribosomal gene expression observed by Bacon et al. (2004); the superscript single prime denotes high (50 % dissolved oxygen tension) and the superscript double prime denotes low (1 % dissolved oxygen tension).

\[
r_{50} = 0.87 \pm 0.23 (12)
\]

Secondly, equation 13 summarizes the data for ribosomal protein genes reported by Kendall et al. (2004); the superscript single prime denotes wild-type \( M. \) tuberculosis and the superscript double prime denotes the \( \text{dosR} \) ‘knockout’ mutant.

\[
r_{50} = 0.93 \pm 0.16 (13)
\]

Thirdly, equation 14 summarizes the data for the control experiment (Voskuil et al., 2004); both the superscript single prime and the superscript double prime refer to the cDNA probes copied from the same RNA fraction. Hence, in the ideal case \( r = 1 \).

\[
r_{50} = 1.00 \pm 0.10 (14)
\]

As expected, it was found (see equations 12, 13 and 14) that the mean expression ratios for ribosomal protein genes are close to unity. Thus, I infer that equation 11b is applicable to expression values obtained when reference and experimental strains grow at the same rate.

Table 3. Ribosomal protein genes studied

<table>
<thead>
<tr>
<th>Tag</th>
<th>Locus</th>
<th>E. coli equivalent</th>
<th>Length (bp)</th>
<th>Tag</th>
<th>Locus</th>
<th>E. coli equivalent</th>
<th>Length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rv0053</td>
<td>rpsF</td>
<td>S6</td>
<td>291</td>
<td>Rv0055</td>
<td>rpsR</td>
<td>S18</td>
<td>255</td>
</tr>
<tr>
<td>Rv0056</td>
<td>rplI</td>
<td>L9</td>
<td>459</td>
<td>Rv0640</td>
<td>rplK</td>
<td>L11</td>
<td>429</td>
</tr>
<tr>
<td>Rv0641</td>
<td>rplA</td>
<td>L1</td>
<td>708</td>
<td>Rv0651</td>
<td>rplF</td>
<td>L10</td>
<td>537</td>
</tr>
<tr>
<td>Rv0652</td>
<td>rplL</td>
<td>L12</td>
<td>393</td>
<td>Rv0682</td>
<td>rpsL</td>
<td>S12</td>
<td>375</td>
</tr>
<tr>
<td>Rv0683</td>
<td>rpsG</td>
<td>S7</td>
<td>471</td>
<td>Rv0700</td>
<td>rpsJ</td>
<td>S10</td>
<td>306</td>
</tr>
<tr>
<td>Rv0701</td>
<td>rplC</td>
<td>L3</td>
<td>654</td>
<td>Rv0702</td>
<td>rplD</td>
<td>L4</td>
<td>672</td>
</tr>
<tr>
<td>Rv0703</td>
<td>rplW</td>
<td>L23</td>
<td>303</td>
<td>Rv0704</td>
<td>rplB</td>
<td>L2</td>
<td>843</td>
</tr>
<tr>
<td>Rv0705</td>
<td>rpsS</td>
<td>S19</td>
<td>282</td>
<td>Rv0706</td>
<td>rplV</td>
<td>L22</td>
<td>594</td>
</tr>
<tr>
<td>Rv0707</td>
<td>rpsC</td>
<td>S3</td>
<td>825</td>
<td>Rv0708</td>
<td>rplP</td>
<td>L16</td>
<td>138</td>
</tr>
<tr>
<td>Rv0709</td>
<td>rpsM</td>
<td>S2</td>
<td>992</td>
<td>Rv0710</td>
<td>rpsQ</td>
<td>S17</td>
<td>411</td>
</tr>
<tr>
<td>Rv0714</td>
<td>rplN</td>
<td>L14</td>
<td>369</td>
<td>Rv0715</td>
<td>rplX</td>
<td>L24</td>
<td>318</td>
</tr>
<tr>
<td>Rv0716</td>
<td>rplE</td>
<td>L5</td>
<td>564</td>
<td>Rv0717</td>
<td>rpsN</td>
<td>S14</td>
<td>486</td>
</tr>
<tr>
<td>Rv0718</td>
<td>rpsH</td>
<td>S8</td>
<td>599</td>
<td>Rv0719</td>
<td>rplF</td>
<td>L6</td>
<td>540</td>
</tr>
<tr>
<td>Rv0720</td>
<td>rplR</td>
<td>L18</td>
<td>369</td>
<td>Rv0721</td>
<td>rpsE</td>
<td>S5</td>
<td>663</td>
</tr>
<tr>
<td>Rv0722</td>
<td>rplD</td>
<td>L30</td>
<td>198</td>
<td>Rv0723</td>
<td>rplO</td>
<td>L30</td>
<td>198</td>
</tr>
<tr>
<td>Rv1015c</td>
<td>rplY</td>
<td>L25</td>
<td>648</td>
<td>Rv1298</td>
<td>rpmE</td>
<td>L31</td>
<td>243</td>
</tr>
<tr>
<td>Rv1630</td>
<td>rpsA</td>
<td>S1</td>
<td>1446</td>
<td>Rv1642</td>
<td>rpmL</td>
<td>L35</td>
<td>195</td>
</tr>
<tr>
<td>Rv1643</td>
<td>rplT</td>
<td>L20</td>
<td>390</td>
<td>Rv2412</td>
<td>rpsT</td>
<td>L30</td>
<td>261</td>
</tr>
<tr>
<td>Rv2441c</td>
<td>rpsA</td>
<td>L27</td>
<td>261</td>
<td>Rv2442c</td>
<td>rplU</td>
<td>L21</td>
<td>315</td>
</tr>
<tr>
<td>Rv2785c</td>
<td>rplO</td>
<td>S15</td>
<td>270</td>
<td>Rv2890c</td>
<td>rpsB</td>
<td>S2</td>
<td>864</td>
</tr>
<tr>
<td>Rv2904c</td>
<td>rplS</td>
<td>L19</td>
<td>342</td>
<td>Rv2909c</td>
<td>rpsP</td>
<td>S16</td>
<td>489</td>
</tr>
<tr>
<td>Rv3442c</td>
<td>rpsI</td>
<td>S9</td>
<td>456</td>
<td>Rv3443c</td>
<td>rplM</td>
<td>L13</td>
<td>444</td>
</tr>
<tr>
<td>Rv3456c</td>
<td>rplQ</td>
<td>L17</td>
<td>543</td>
<td>Rv3458c</td>
<td>rpsD</td>
<td>S4</td>
<td>606</td>
</tr>
<tr>
<td>Rv3459c</td>
<td>rpsK</td>
<td>S11</td>
<td>420</td>
<td>Rv3460c</td>
<td>rpmF</td>
<td>L36</td>
<td>345</td>
</tr>
<tr>
<td>Rv3688c</td>
<td>rpsU</td>
<td>S21</td>
<td>465</td>
<td>Rv3924c</td>
<td>rpmH</td>
<td>L34</td>
<td>144</td>
</tr>
</tbody>
</table>

Table 4. How many of these genes encode ribosomal proteins?

<table>
<thead>
<tr>
<th>Gene tag</th>
<th>Gene locus</th>
<th>E. coli equiv.</th>
<th>Essential gene</th>
<th>Size of protein (aa)</th>
<th>No. of amino acid similarities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rv0717</td>
<td>rpsN1</td>
<td>S14</td>
<td>?</td>
<td>61</td>
<td>36</td>
</tr>
<tr>
<td>Rv2056c</td>
<td>rpsN2</td>
<td>S14</td>
<td>No</td>
<td>101</td>
<td>40</td>
</tr>
<tr>
<td>Rv0055</td>
<td>rpsR1</td>
<td>S18</td>
<td>?</td>
<td>84</td>
<td>54</td>
</tr>
<tr>
<td>Rv2055c</td>
<td>rpsR2</td>
<td>S18</td>
<td>No</td>
<td>88</td>
<td>54</td>
</tr>
<tr>
<td>Rv0105c</td>
<td>rpmB1</td>
<td>L28</td>
<td>No</td>
<td>94</td>
<td>53</td>
</tr>
<tr>
<td>Rv2058c</td>
<td>rpmB2</td>
<td>L28</td>
<td>No</td>
<td>78</td>
<td>53</td>
</tr>
<tr>
<td>Rv2057c</td>
<td>rpmG1</td>
<td>L33</td>
<td>No</td>
<td>54</td>
<td>30</td>
</tr>
<tr>
<td>Rv0634B</td>
<td>rpmG2</td>
<td>L33</td>
<td>?</td>
<td>55</td>
<td>55</td>
</tr>
</tbody>
</table>
B. Reference and experimental strains grown at different rates

The data analysed were obtained from samples collected from a culture of M. tuberculosis (Voskuil et al., 2004) over a period of 60 days. Cells collected at mid-exponential phase were used as the reference strain. Results are presented (see Tables 5, 6 and 7 and Fig. 2) as plots of expression ratios versus the periods of cell culture (expression profiles). Wherever possible the precision of the microarray data was improved by obtaining average expression ratios \( \langle r_{(i)} \rangle \pm SD \) for sets of genes encoding proteins which have related functions; for example, ribosomal proteins (see Fig. 2).

Two expression profiles (I and II) are identified in Fig. 2 in order to denote proteins with the properties of \( p_{(i)} \) and \( p_{(k)} \) (see Table 2). By definition (see equation 11c, Table 2), the copy numbers \( n_{c-p(j)} \) of protein of the \( p_{(j)} \) family are independent of \( \mu \) \( (n'_{c-p(j)}/n'_{c-p(j)}=1) \) and (see equation 11e, Table 2) the copy numbers of protein of the \( p_{(k)} \) family are directly proportional to \( n_{R(\text{av})} \) \( (n'_{c-p(k)}/n'_{c-p(k)}=n'_{R(\text{av})}/n'_{R(\text{av})}) \). The specific growth rate is expected to decrease after exponential phase as the supply of nutrients diminishes. The slower growth rate is reflected in Fig. 2(a)–(i) as a decrease in expression ratios with the period of cell culture. However, the expression ratios were found to be similar in values for days 6–8, possibly because the growth rate remained constant over this period due to adaptation to an alternative substrate in the complex medium used.

Proteins involved in DNA partition and synthesis. The mycobacterial cell replicates its genome once during each cell division cycle and then partitions the two genomes equally between the two daughter cells. I propose that proteins involved in DNA partition are required to the same extent by all cells irrespective of growth rate. This proposal is supported by our knowledge of the three-dimensional structure and function of FtsK (Massey et al., 2006). In other words, for a particular strain the copy number \( n_{c-p(j)} \) of ‘partition’ proteins, \( p_{(j)} \), is constant at all growth rates so that \( n''_{c-p(j)}/n'_{c-p(j)}=1; \) see equation 11c (Table 2).

The average \( r_{(3)} \) for three genes encoding proteins involved in DNA partition, \( \text{ftsK}, \text{parA} \) and \( \text{parB} \), at different periods of cell culture are shown in Fig. 2(a). This profile labelled I is used to represent \( r_{(j)} \) versus the period of cell culture. The \( r_{(5)} \) for five proteins involved in DNA replication lie close to profile I (see Fig. 2b), suggesting that the numbers of copies per population-average cell of these proteins also remain constant and independent of growth rate. It was noted (see Fig. 2c) that \( r_{(5)} \) for five proteins involved in DNA partition, \( \text{dnaE1}, \text{dnaE2}, \) and \( \text{dnaE3} \), followed the pattern obtained for ribosomal proteins which is the reference profile for \( r_{(k)} \) (see curve II, Fig. 2d), while the data for \( \text{dnaE2} \) were too erratic to be helpful. Thus, the possibility remains that, as the growth rate diminishes, the availability of the \( x \)-chain of DNA polymerase III governs the concentration of DNA polymerase III holoenzyme.

Proteins involved in protein biosynthesis. The average value of \( r \) (\( r_{(50)} \)) for the set of 50 ribosomal proteins was found to diminish with increasing periods of cell culture. Profile II is the plot of the average values of \( r_{(50)} \) against the period of cell culture (Fig. 2d) and is equivalent to the plot of \( r_{(k)} \) against time (see Table 2). Eight factors required for protein synthesis including elongation factor EF-Tu (which participates in the rate-limiting step in peptide bond...
Table 5. Representative microarray data obtained for *M. tuberculosis* grown in batch culture: profiles of expression ratios of highly upregulated and other selected genes

<table>
<thead>
<tr>
<th>Tag</th>
<th>Locus</th>
<th>Essential†</th>
<th>Parameter*</th>
<th>Period (days) of cell culture</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0‡</td>
</tr>
<tr>
<td>Rv0467</td>
<td>icl</td>
<td>?</td>
<td>r</td>
<td>0.91</td>
</tr>
<tr>
<td>Rv2031c</td>
<td>hspX</td>
<td>No</td>
<td>n′_c-p(i)/n′_c-p(i)</td>
<td>0.90</td>
</tr>
<tr>
<td>Rv2150c</td>
<td>ftsZ</td>
<td>Yes</td>
<td>r</td>
<td>ND</td>
</tr>
<tr>
<td>Rv2626c</td>
<td>–</td>
<td>?</td>
<td>n′_c-p(i)/n′_c-p(i)</td>
<td>0.99</td>
</tr>
<tr>
<td>Rv2660c</td>
<td>–</td>
<td>?</td>
<td>r</td>
<td>1.51</td>
</tr>
<tr>
<td>Rv3332c</td>
<td>dosR (devR)</td>
<td>No</td>
<td>n′_c-p(i)/n′_c-p(i)</td>
<td>0.86</td>
</tr>
</tbody>
</table>

†Not reported in Tuberculist gene list; ND, no result reported.

*Expression ratios (r) were reported by Voskuil et al. (2004); for details see legend to Fig. 2. Standard deviations are omitted to simplify the presentation. Values of n′_c-p(i)/n′_c-p(i) (shown in bold italics) were calculated as described in the text.

‡Determined by Himar1-based transposon mutagenesis in strain H37Rv (Sassetti et al., 2003).

§Mid-exponential-phase cells: DNA′ and cDNA′ were copied from samples of the same RNA preparation.

...formation) form a group. Their average r-value (r_(8)) follows the trend observed for ribosomal proteins; that is, within experimental error, values of r_(8) follow curve II (see Fig. 2d), as shown in Fig. 2(e). It is inferred that the number of copies of these eight factors per ribosome remains constant irrespective of growth rate. The expression of genes encoding three factors, fusA2, prfA and infB, was found to vary with the period of incubation in a different way, as judged by the profile of r_(3) values (see Fig. 2f), which is similar to profile I (see Fig. 2a). It is concluded that the copy numbers of these proteins per population-average cell are independent of growth rate.

In contrast with fusA1 (Sassetti et al., 2003), fusA2 is a non-essential gene (Lamichhane et al., 2003). Both FusA1 and FusA2 are classified as having elongation factor EF-G activity. However, they differ significantly in their amino acid sequences; they share 221/701 identical amino acids, which is comparable with the 233 identical amino acids shared between FUS1 and elongation factor EF-G of *Escherichia coli*. Both infB, encoding initiation factor IF-2, and prfA, encoding peptide chain release factor RF-1, are essential genes (Sassetti et al., 2003).

Table 6. Estimates of m′_p*(av)/m′_p*(av)

(n′_c-p(i)/n′_c-p(i))_av is considered to be an approximation for m′_p*(av)/m′_p*(av) (see footnote ‡ of Table 2).

<table>
<thead>
<tr>
<th>Period of cell culture (days)</th>
<th>r_(i)</th>
<th>r_(i)</th>
<th>(n′_c-p(i)/n′_c-p(i))_av</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(A)</td>
<td>(B)</td>
<td>(A/B)</td>
</tr>
<tr>
<td>0</td>
<td>0.99</td>
<td>0.95</td>
<td>1.04</td>
</tr>
<tr>
<td>6</td>
<td>0.74</td>
<td>0.82</td>
<td>0.93</td>
</tr>
<tr>
<td>8</td>
<td>0.75</td>
<td>0.51</td>
<td>0.92</td>
</tr>
<tr>
<td>14</td>
<td>0.54</td>
<td>0.47</td>
<td>1.06</td>
</tr>
<tr>
<td>24</td>
<td>0.42</td>
<td>0.41</td>
<td>1.07</td>
</tr>
</tbody>
</table>

‡v is the number of ORFs involved in the summation; in this example; v=206 (ORFs Rv0001–Rv0206).

Table 7. Changes in the properties of population-average cells of *M. tuberculosis* grown in batch culture derived from microarray measurements

<table>
<thead>
<tr>
<th>Variable</th>
<th>Eqn</th>
<th>Time (days) of cell culture</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>m′_p*(av)/m′_p*(av)</td>
<td>11h</td>
<td>1.00</td>
</tr>
<tr>
<td>m′_R*(av)/m′_R*(av)</td>
<td>11d</td>
<td>1.04</td>
</tr>
<tr>
<td>β′_R</td>
<td>11g</td>
<td>1.00</td>
</tr>
<tr>
<td>m′_RNA*(av)/m′_RNA*(av)</td>
<td>11g</td>
<td>1.00</td>
</tr>
<tr>
<td>β′_R·n′_R*(av)/β′_R·n′_R*(av)</td>
<td>†</td>
<td>1.00</td>
</tr>
</tbody>
</table>

†This expression is the product of β′_R/n′_R and m′_RNA*(av)/m′_RNA*(av), which is the ratio of the number of ribosomes actively synthesizing protein.
A. Genes encoding proteins needed for DNA partition and synthesis

B. Genes encoding ribosomal proteins and factors needed for protein synthesis

C. Genes encoding proteins needed for RNA synthesis

D. Genes encoding proteins whose copy numbers increase twofold at some stage
Proteins involved in RNA synthesis. tRNA constitutes a major part of the RNA fraction (mRNA_{av}) of population-average cells. The synthesis of tRNA is the rate-limiting step in ribosome synthesis (for a review see Keener & Nomura, 1996). Thus, the rate of synthesis of proteins involved in tRNA synthesis might be expected to coordinate with the rate of synthesis of ribosomal proteins. As anticipated, the r-values (r_{av}) of six genes encoding proteins involved in RNA replication were found to follow profile II (Fig. 2a). Two genes, nusA and nusB, encoding proteins implicated in antitermination activity during the course of tRNA synthesis followed a different pattern (see Fig. 2h; the observed r-values followed profile I, Fig. 2a, rather than profile II, Fig. 2d). It is inferred that the copy numbers of NusA and NusB per population-average cell vary little with growth rate.

The tRNA pool functions in protein biosynthesis through the formation of aminoacyl-tRNA derivatives. The rate-limiting step in peptide chain formation is the interaction of a ternary complex of aminoacyl-tRNA, elongation factor EF-Tu and GTP with the A-site of the ribosome. Aminoacyl-tRNA synthases are responsible for charging tRNA. The r-values (r_{av}) of 16 aminoacyl-tRNA synthase genes were examined (see Fig. 2i). The dependence of r_{av} on the period of cell culture was found to follow profile II (Fig. 2i) for the first 14 days, indicating that the number of copies of aminoacyl-tRNA synthases per ribosome is constant over this period. The expression ratios found after longer periods of cell culture are greater than those expected for profile II, suggesting that the number of copies per ribosome increases at very slow growth rates.

Expression profiles other than profile I and profile II were observed (see Fig. 2j–l). Each profile is consistent with a doubling of the copy number (n_{c,p(i)} = 2n_{c,p(i)}) of the encoded protein at some stage during the 60 day period of cell culture. Other examples of periodically expressed genes are presented in Table 5. Large changes in the expression ratios were found for hspX, Rv2626c and Rv2660c. The non-essential gene dosR was found to be upregulated after 14 days of cell culture. The expression ratio of icl diminished as the period of cell culture increased. The expression ratio for the essential gene ftsZ was found to follow profile II, indicating that the number of copies of FtsZ per cell is proportional to the number of ribosomes per cell; that is, FtsZ is a member of the p_{av} family of proteins.

C. Evaluation of parameters commonly used to describe cell growth

The growth of a cell culture is often described by the specific growth rate (μ) and a measure of macromolecular composition such as the ratio DNA : RNA : protein or more recently by μ and the macromolecular composition of a culture average cell (see the Theory section).

---

**Fig. 2.** Representative microarray data obtained for *M. tuberculosis* grown in batch culture. Profiles of expression ratios of ORFs encoding proteins engaged in A, DNA synthesis and segregation; B, protein synthesis; C, RNA synthesis; and D, proteins whose copy numbers (n_{c,p(i)}) increase twofold at some stage during culture. Expression ratios were obtained from Voskuil et al. (2004) for a culture of *M. tuberculosis* clinical isolate 1254 grown at 37 °C in 7H9 medium (supplemented with BSA, NaCl, glucose and glycerol) with shaking at 90 r.p.m. Cell culture was continued for 60 days and samples were removed at the times specified in the figure. Time 0 days refers to a mid-exponential-phase culture of OD_{600} 0.15. The expression ratio at this time point was obtained by using the RNA isolated to prepare one sample of cDNA labelled with fluorophor F’ and a second sample labelled with fluorophor F. In each case the reference strain is the culture of mid-exponential-phase cells. The experimental strain is the culture at the time point specified. The expression profiles reflect the changes in gene expression in response to depletion of nutrients in the culture medium. r_{av}, etc., denote the average expression ratio of the number of related genes specified in the subscript. Standard deviations are shown by the error bars. Profile I is defined in (a) and is equivalent to the profile of r_{av} versus time. ORF_{av} encodes protein p_{av}, whose copy number n_{c,p(i)} is independent of growth rate. Profile II is defined in (d) and is equivalent to the profile of r_{av} versus time. ORF_{av} encodes protein p_{av}; n_{c,p(i)} is directly proportional to r_{av}. Profile III (a broken line) is specified in (l) and is equivalent to the profile of r_{av} versus time when ORF_{av} encodes a protein which, at some time between 0 and 6 days, increases twofold in copy number and then remains at this value as cell culture proceeds. A, Genes encoding proteins needed for DNA partition and synthesis. (a) ORFs encoding proteins implicated in DNA partition. r_{av} is the average expression ratio for parA, parB and ftsK. (b) ORFs encoding proteins needed for DNA synthesis. r_{av} is the average expression ratio for dnaN, dnaG, polA, dnaQ and dnaZX. (c) Expression profile for dnaE1. B, Genes encoding ribosomal proteins and factors needed for protein synthesis. (d) ORFs encoding ribosomal proteins. r_{av} denotes the average expression ratio for the 50 ORFs listed in Table 3 encoding ribosomal proteins. (e) ORFs encoding protein factors (Group I) needed for protein synthesis. r_{av} denotes the average expression values for fusA1, fusA2, prfA and infB. C, Genes encoding proteins needed for RNA synthesis. (g) ORFs encoding proteins needed for RNA synthesis. r_{av} (open circles). (i) ORFs encoding rRNA synthases. r_{av} denotes the average expression ratio for, nusG, rpoB, rpmC, rho, rpmA and gpaSL. (h) Expression ratios for nusA (filled circles) and nusB (open circles). (j) Genes encoding proteins whose copy numbers increase twofold at some stage during cell culture. (j), (k) and (l) respectively show expression ratios for Rv0165c, Rv3336c, Rv3580c and Rv3598c. D, Genes encoding proteins whose copy numbers increase twofold at some stage during cell culture. (j), (k) and (l) respectively show expression ratios for Rv0165c, Rv3336c, Rv3580c and Rv3598c.
Relative values of $m_{p(avg)}$ were deduced in the following way. At time $t$ the ratio $r_{(i)}(t)/r_{(j)}(t)$ provides the ratio $n'_c-p_{(i)}/n'_c-p_{(j)}$. Thus the sum of $\sum i=1$ to $i=n_{ORF/g}$ divided by $n_{ORF/g}$ provides an average value $(n'_c-p_{(i)}/n'_c-p_{(j)})_{av}$ which should reflect the ratio $m'_{p(avg)}/m'_{p(avg)}$. For example, if after 60 days of cell culture $m'_{p(avg)}$ is 0.5, $m'_{p(avg)}$ then $(n'_c-p_{(j)})_{av}$ would also tend towards 0.5. In practice, it was found that the sum of $\sum (r_{(i)})_{av}$ from $i=1$ to $i=206$ provided reliable estimates of $(n'_c-p_{(i)}/n'_c-p_{(j)})_{av}$. The calculated values (see Table 6) suggest that although the composition of the protein fractions may vary with growth rate the mass of protein per cell $m'_{p(avg)}$ is close to $m'_{p(avg)}$ for all periods of cell culture.

Inspection of the sample of 206 genes revealed that in 74 cases the number of copies of protein $p_{(i)}$ per cell, $n_{p_{(i)}}$, changed very little ($n'_c-p_{(i)} \approx n'_c-p_{(i)}$); in 44 cases $n'_c-p_{(i)}$ was found to increase with the period of cell culture; in 86 cases $n'_c-p_{(i)}$ was found to decrease with the period of culture. However, changes in the ratio $n'_c-p_{(i)}/n'_c-p_{(i)}$ were relatively small (<2) and usually less than the corresponding decrease noted for ribosomal proteins (see Fig. 2d). Increases in the number of copies of protein $p_{(i)}$ per cell with the period of cell culture followed different patterns; examples are presented in Fig. 2(j–l) and Table 5. Two-dimensional gel analysis of the overall pattern of cell proteins indicates very little change in the number of proteins identified with the period of cell culture, although the relative abundances of individual proteins (as revealed by the intensities of staining) may vary (Bets et al., 2002).

The equations summarized in Table 2 allow evaluation of the relative values of $\mu$, $m_{RNA(avg)}$ and $m_{p(avg)}$. Values of $r_{(i)}(j)$ and $m'_{p(avg)}/m'_{p(avg)}$ are needed to solve the appropriate equations. Profile I (Fig. 2a) is considered to provide values of $r_{(i)}(j)$ as a function of the period of cell culture; that is, the copy numbers of proteins $p_{(i)}$ (DNA partition proteins) are regarded as independent of growth rate: $n'_c-p_{(i)}/n'_c-p_{(j)}=1$. Profile II (Fig. 2d) observed for ribosomal proteins has the properties of $r_{(i)}(k)$ as a function of the period of cell culture. In this case equation 11a reduces to equation 11c provided that the proportion of $m_{RNA(avg)}$ that is rRNA remains unchanged. Then, $n'_c-p_{(k)}/n'_c-p_{(k)}$ is equal to $n'_R_{(av)}/n'_R_{(av)}$, which is equal to $m'_{RNA(avg)}/m'_{RNA(avg)}$. The ratio $m'_{p(avg)}/m'_{p(avg)}$ was estimated as a function of the period of cell culture (see equation 11d), as summarized in Table 6. These relationships were used to calculate the parameters presented in Table 7.

Maximum values of $\mu'/\mu'$ (see Table 7) were obtained by means of equation 11e (Table 2) on the basis of the assumption that $c'_{a_{(a)k}}(k)=c'_{a_{(a)k}}(k)$. Maximum values of $\mu'$ were then deduced by supposing that $\mu'=0.04 h^{-1}$ ($t_{f}=16 h$). The following values were obtained: at day 6 and at day 8 $\mu'=0.028 h^{-1}$ ($t_{f}=24 h$); at day 14 $\mu'=0.015 h^{-1}$ ($t_{f}>46 h$); at day 24 $\mu'<0.008 h^{-1}$ ($t_{f}>80 h$); at day 60 $\mu'<0.004 h^{-1}$ ($t_{f}>160 h$). Technically it is very difficult to measure $\mu$ during stationary phase. However, it is to be expected that $\mu$ will diminish as the nutrient supply is reduced.

After 60 days of cell culture the cellular RNA content $m_{RNA(avg)}$ and the ribosome content $n_{RNA(avg)}$ were estimated to have decreased to approximately one-quarter of the content of mid-exponential-phase cells. This decrease in $m_{RNA(avg)}$ and $n_{RNA(avg)}$ is consistent with data reported by Beste et al. (2005) for Mycobacterium bovis BCG, the vaccine strain of M. tuberculosis. These authors grew M. bovis BCG in a chemostat at two different rates, $\mu=0.03 h^{-1}$ ($t_{f}=23 h$) and $\mu=0.01 h^{-1}$ ($t_{f}=69 h$), in order to represent mid-exponential-phase and stationary-phase cells. Their data reveal that, allowing 1.4 genome equivalents per cell, the slower-growing cells had 22% of the RNA content of the faster-growing cells.

The fraction of cells actively synthesizing protein was also found to diminish over 60 days of cell culture (see Table 7) to no more than 38% of the value for mid-exponential-phase cells. In other words at day 60 more than half of the ribosomes were not involved in synthesizing protein. Stationary-phase cells of E. coli also maintain an excess of unprogrammed (non-translating) ribosomes (in this case in the form of 100S dimers of 70S ribosomes) which allow the cell to respond rapidly on relief of nutritional constraints (McCarthy, 1960; Feiss & DeMoss, 1965; Wada et al., 1990; for a review see Keener & Nomura, 1996).

In summary, the information presented in Table 7 provides both direct and indirect support for the theoretical framework which is summarized by the equations shown in Table 2.

D. Review of the support for equation 11 (see Table 2)

The data presented in section A (see Fig. 1) confirm the proposal that equation 11a is reduced to equation 11b when reference and experimental strains grow at the same specific growth rate. In this case the expression ratio is directly proportional to the ratio of the copy numbers $n'_c-p_{(i)}/n'_c-p_{(i)}$ of the encoded protein, as shown for ribosomal proteins.

A notable feature of section B is the number of different expression profiles that were found (see Fig. 2 and Table 5). This diversity indicates that the expression ratio reflects the function of the encoded protein in cell metabolism. The equations listed in Table 2 provide the theoretical basis for analysing an individual profile. The reference profiles I and II respectively characterize genes encoding the $p_{(i)}$ and $p_{(k)}$ families of genes. Eight genes, including five that encode proteins needed for DNA partition and synthesis, are represented by profile I. Sixty-four genes encoding ribosomal proteins, protein factors needed for protein synthesis and proteins needed for RNA synthesis are all represented by profile II. The equations underlying profiles I and II (see Table 2) were used to calculate changes in the
properties of population-average cells during the course of cell culture (see Table 7). The trends reported in Table 7 are in accord with our knowledge of the growth of bacterial cells in batch culture. This outcome provides support for equation 11a, etc. Table 7 is based on experimental data which comprises expression ratios only.

A contrasting approach is that of Nie et al. (2006), who used multiple regression analysis to correlate measurements of mRNA abundance and protein abundance in Desulfovibrio vulgaris. They found a modest correlation between these two quantities, which is largely explained by the technical problems they encountered. Microarray analysis was used to measure mRNA abundances. Tryptic digests of the protein fraction were separated by high-performance liquid chromatography and fractions were monitored by tandem mass spectrometry. No indication was given as to how these measurements might relate to other cell properties.

**Concluding remarks**

The results presented above complement the studies of Bacon et al. (2004), Betts et al. (2002), Kendall et al. (2004) and Voskuil et al. (2004) by extending the scope of the analysis of microarray data.

Many techniques, including microarray analysis, provide information about a population of cells. For this reason the concept of a population-average cell is both accurate and helpful. This concept provides the basis for developing a virtual schematic cell that summarizes the properties deduced from studies of cell cultures. In turn, the schematic cell can be an aid to designing new experiments.

In this report I have used the concept of a population-average cell to show that the expression ratio measured by microarray analysis is dependent not only on relative copy numbers of the encoded protein but also on several other cell parameters. However, when reference (for example wild-type) and experimental (for example mutant) strains grow at the same rate the expression ratio is equal to the ratio of the copy numbers of the encoded protein (the other terms cancel out).

The analysis of microarray data obtained when experimental and reference strains grew at different rates led to the recognition that in two cases the expression ratio is independent of the copy numbers of the encoded protein: namely genes encoding proteins \( p_{ij} \) whose copy numbers vary little with growth rate, and genes encoding proteins \( p_{1} \) whose copy numbers are directly proportional to the number of ribosomes per cell. Proteins of the \( p_{ij} \) family are potentially useful as cell markers; for example, fluorescently labelled protein could be used to monitor both cell size and cell numbers.

At present few data are available for copy numbers of proteins of the tubercule bacillus. Thus, the theoretical framework could prove helpful in the study of metabolic pathways because of the possibility of monitoring changes with growth rate in the copy numbers of proteins involved in a particular process or pathway. This proposal is illustrated by three cell processes that were analysed in this report: DNA replication, protein synthesis and RNA synthesis. This study also highlights the extra information that could be obtained by applying microarray studies to culture of known growth rates and macromolecular compositions. Further understanding of metabolic processes gained in this way should lead to the identification of new drug targets and so facilitate drug discovery. Finally, the theoretical framework was developed for bacteria in general and its application is not restricted to M. tuberculosis.

**APPENDIX**

**Correlation between the expression ratio \( r_0 \) and the number of copies \( n_{c-p(i)} \) of protein \( p_{ij} \) per cell**

**Transcription/translation coupling.** Coupling between the processes of bacterial transcription and translation has long been accepted (Stent, 1964; Byrne et al., 1964; Miller et al., 1970). Transcription/translation coupling is thought to require that the rates of mRNA elongation \( e_{molRNA} \) nucleotides h\(^{-1}\) and peptide chain elongation \( e_{aa} \) amino acid residues h\(^{-1}\) are co-ordinated so that \( e_{molRNA}=3e_{aa} \) where the factor 3 reflects the number of nucleotides per codon. When ribosomes are translating a previously synthesized mRNA the rate-limiting step in protein biosynthesis is the rate of interaction of tc (a ternary complex formed between aminocycltRNA, elongation factor EF-Tu and GTP) with the A-site of a ribosome (Pape et al., 1998). In order for transcription and translation to be coupled the rate-limiting step \( d(tc-A-site)/dt \) amino acids h\(^{-1}\) must be equal to or faster than one-third of the elongation rate \( e_{molRNA} \) nucleotides h\(^{-1}\) of nascent mRNA\( (i) \). Suppose that the maximum rate for peptide chain elongation is \( e_{aa(i)} \) which is achieved when the formation of the peptide bond is rate limiting. Then, \( e_{molRNA(i)}=3e_{aa(i)} \). When \( e_{molRNA(i)}<3e_{aa(i)} \) the rate of transcription, by limiting the availability of codons, becomes the rate-limiting step in the synthesis of protein \( p_{ij} \).

Transcriptional control processes determine the rate of transcription of ORF\( (i) \). The maximum number of RNAPs per ORF\( (i) \) is determined by the size \( \psi \) bp of the region of DNA involved in the formation of an open complex RNAP\(_{oc} \) with RNAP to initiate transcription. The need for RNAP\(_{oc} \) to move clear of the promoter restricts the maximum number of RNAPs per ORF\( (i) \) to one per \( \psi \) base pair. DNase protection studies (Ozoline & Tsaganov, 1995) show that \( \psi \) ranges from 60 bp to approximately 80 bp for promoters that are active without a regulator and from 70 bp to 100 bp for regulated promoters.

The process of translation is thought to protect mRNA from degradation by RNases (degradosomes) (Grunberg-Manago, 1999; Regnier & Arnaudo, 2000; Leroy et al., 2002; Khodursky & Bernstein, 2003). It is proposed that the number, \( \sigma \), of nucleotides protected by translation is related to the size of a region of mRNA protected by a ribosome and the number of nucleotides that form a binding site for a degradosome. The largest dimension of the bacterial 70S ribosome is 22.5 + 25 nm (Noller & Nomura, 1996), which is equivalent in length to a sequence of 60–75 nucleotides, allowing 0.34 nm (3.4 Å) per nucleotide. The maximum value of \( \sigma \) corresponds to the length of the region protected by ribosomes plus a region smaller than the number of nucleotides that form a binding site for a degradosome. For example, unprotected stretches of mRNA of 20 nucleotides or more were considered to be capable of binding degradosomes and, hence,
to be very rapidly degraded. Thus, a sequence of two regions of 10 nucleotides separated by a ribosome-binding site may approach the limiting size for a sequence protected from RNase action; namely a sequence of 80–95 nucleotides. Kinetic studies of the mRNA fraction of *E. coli* summarized by Bremer & Dennis (1996) led to estimates for *σ* of 55–80 nucleotides per ribosome depending on the growth rate. Thus, a value of *σ* = 80 ± 20 nucleotides per ribosome summarizes the several estimates.

**Properties of a specified ORF (ORF(i)).** The properties of an ORF that is actively being transcribed can be related to the population-average cell as described below. A particular open reading frame ORF(i) encodes a particular protein *p(i)*, which is present as *n_p(i)* copies of *p(i)* per population-average cell; *n_p(i)* ≥ 0 copies per cell. The length of *p(i)* is defined as *l_{nuc}(i)* amino acid residues so that the mass *m_p(i)* of *p(i)* is the product of *l_{nuc}(i)* and *m_{nuc}* (the average mass of an amino acid residue). Hence, *m_p(i)* is the sum from *i = 1* to *i = n_{ORFs(i)}* of the terms *m_{nuc}* *n_p(i)* *l_{nuc}(i)* (see equation 1A). The limits of the summation are omitted to simplify the presentation; *n_{ORFs(i)}* is the number of ORFs per genome.

\[
m_{p(i)} = m_{\text{nuc}} \sum (n_p(i) l_{\text{nuc}}(i))
\]  

(1A)

Equation 1A allows \( n_p(i) \), the specific protein synthesis rate for a population-average cell, to be expressed in terms of a specified ORF. For exponential growth the term \( n_p(i) \) is defined by equation 2A, where \( \mu \) h \(^{-1}\) is the specific growth rate:

\[
\phi_p(i) = m_{p(i)} \mu
\]  

(2A)

Substitution for \( m_{p(i)} \) in equation 2A leads to equation 3A:

\[
\phi_p(i) = m_{\text{nuc}} \mu \sum (n_p(i) l_{\text{nuc}}(i))
\]  

(3A)

Thus, the specific protein synthesis rate \( \phi_p(i) \) is equal to the sum of the specific protein synthesis rates \( \phi_{p(i)} \). The parameter \( \phi_{p(i)} \) is thus defined by equation 4A:

\[
\phi_{p(i)} = m_{\text{nuc}} \mu \sum n_p(i) l_{\text{nuc}}(i)
\]  

(4A)

The rate \( \phi_{p(i)} \) (see Fig. 1) is also equal to the product of *n_{R(i)}*, the number of ribosomes translating transcripts of ORF(i) and \( \phi_{e(i)} \), the rate (amino acids h \(^{-1}\)) at which a ribosome elongates a peptide chain; see equation 5A:

\[
\phi_{p(i)} = m_{\text{nuc}} \phi_{e(i)} c_{p(i)} e_{\text{sat}(i)}
\]  

(5A)

The number of ribosomes per population-average cell directly involved in protein synthesis is the product \( \beta_R \phi_{p(i)} \) where \( \beta_R \) is the fraction of ribosomes programmed with mRNA and \( n_{R(i)} \) is the number of ribosomes per population-average cell. The parameter \( n_{R(i)} \) can also be expressed as a summation; see equation 6A:

\[
\beta_R = (1/\beta_R) \sum n_{R(i)}
\]  

(6A)

The parameter \( n_{R(i)} \) can be related to \( \mu \) and \( e_{\text{sat}(i)} \) by equating the right-hand sides of equations 4A and 5A, leading to equation 7A:

\[
n_{R(i)} = n_p(i) e_{\text{sat}(i)} (\mu/e_{\text{sat}(i)})
\]  

(7A)

The RNA fraction comprises three principal components, namely rRNA, tRNA and mRNA. In *E. coli* the proportions are 83 % rRNA, 15 % tRNA and 2 % mRNA (Bremer & Dennis, 1996). Thus, \( m_{\text{RNA(i)}} \) is conveniently expressed as in equation 8A, as the sum of the masses of the individual components:

\[
m_{\text{RNA(i)}} = m_{\text{mRNA(i)}} + m_{\text{tRNA(i)}} + m_{\text{rRNA(i)}}
\]  

(8A)

The parameter \( m_{\text{mRNA(i)}} \) can be expressed as the sum of the masses of all of the transcripts of all of the ORFs, \( \sum m_{\text{mRNA(i)}} \), where

\[
m_{\text{mRNA(i)}} \geq 0
\] depending on whether or not ORF(i) is expressed; see equation 9A:

\[
m_{\text{mRNA(i)}} = \sum m_{\text{mRNA(i)}}
\]  

(9A)

The parameter \( m_{\text{mRNA(i)}} \) can be evaluated on the basis of the assumption that RNases rapidly degrade mRNA that is unprotected by ribosomes. Suppose that each ribosome protects \( \sigma \) nucleotides from RNase action. Then \( n_{R(i)} \) ribosomes translating mRNA(i) will protect \( \sigma n_{R(i)} \) nucleotides and mRNA(i) is specified by equation 10A, where \( m_{\text{nuc}} \) is the average mass of a nucleotide:

\[
m_{\text{mRNA(i)}} = m_{\text{nuc}} \sigma n_{R(i)}
\]  

(10A)

Substituting for \( n_{R(i)} \) in equation 10A leads to equation 11A:

\[
m_{\text{mRNA(i)}} = m_{\text{nuc}} \sigma n_p(i) l_{\text{nuc}}(i) (\mu/e_{\text{sat}(i)})
\]  

(11A)

Equation 11A may be presented in an alternative form (see equation 12A) by defining a new parameter \( n_{R(i)} \) which is also a measure of the transcriptional activity of ORF(i). The number of mRNA(i) nucleotides per ORF(i) is \( m_{\text{mRNA(i)}}/m_{\text{nuc}} \) and the number of nucleotides per full-length transcript of the coding region \( 3(l_{\text{nuc}}(i) + 1) \), there being three nucleotides per codon and one termination codon per ORF. The approximation that \( 3(l_{\text{nuc}}(i) + 1) \) is used throughout without a significant loss of accuracy. Hence, \( n_{R(i)} \), the number of transcript equivalents, is defined by equation 12A:

\[
n_{R(i)} = m_{\text{mRNA(i)}}/(m_{\text{nuc}} 3 l_{\text{nuc}}(i))
\]  

(12A)

Hence, equation 11A may be rearranged to make \( n_{R(i)} \) the subject after each side of the equation is multiplied by \( 3 \):

\[
n_{R(i)} = m_{\text{mRNA(i)}}/(3 m_{\text{nuc}} l_{\text{nuc}}(i)) = n_p(i) (\sigma/3) (\mu/e_{\text{sat}(i)})
\]  

(13A)

Equations 11A, 12A and 13A need to be modified in order to reflect current experimental procedures. For example, in qRT-PCR experiments the quantity measured is regarded as the number of transcripts of ORF(i) per unit mass of RNA. Effectively, the measured quantity is \( n_{R(i)}/(m_{\text{RNA(i)}}) \); this quotient is defined by dividing both sides of equation 13A by \( m_{\text{RNA(i)}} \) (see equation 14A):

\[
n_{R(i)}/(m_{\text{RNA(i)}}) = n_p(i) (\sigma/3) (\mu/e_{\text{sat}(i)})/m_{\text{RNA(i)}}
\]  

(14A)

Note that equations 13A and 14A provide the basis for the quantitative analysis of qRT-PCR measurements. The parameter measured in microarray experiments is the ratio *r* of two appropriately fluorescently labelled cDNA species competing for binding sites presented by DNA immobilized on a solid surface (see Fig. 1).

Suppose that prime and double prime superscripts denote competing cDNA species and that the two cDNA species are prepared from equivalent masses of the RNA substrates. Hence, *r* is defined by equations (15A) and (16A):

\[
r(i) = (m_{+\text{RNA}(i)}/m_{\text{RNA}(i)}) \times (m_{-\text{RNA}(i)}/m_{\text{RNA}(i)})
\]  

(15A)

\[
r(i) = (n_{+\text{R(i)}/n_{-\text{R(i)}}}) \times (m_{+\text{RNA}(i)}/m_{-\text{RNA}(i)})
\]  

(16A)

Equations (15A) and (16A) are equivalent because substitution for either \( m_{\text{RNA(i)}} \) or \( n_{R(i)} \) leads to equation 17A:

\[
r(i) = (n_{+\text{R(i)}/n_{-\text{R(i)}}}) \times (\mu/3) \times (\mu/e_{\text{sat}(i)}) \times (m_{+\text{RNA}(i)}/m_{-\text{RNA}(i)})
\]  

(17A)

Equations (15A–17A) link the measured quantity *r* (see Fig. 1) with four cell parameters, namely \( n_{+\text{R(i)}} \), the number of copies of protein \( p(i) \) per population-average cell; \( \mu \), the specific growth rate; \( e_{\text{sat}(i)} \), the polypeptide chain elongation rate; and mRNA(i).
ACKNOWLEDGEMENTS

I thank Simon A. Cox for his help in the preparation of this manuscript and my colleagues Dr Michael G. Sargent and Dra Maria J. Garcia of the Universidad Autonoma de Madrid for their constructive comments.

REFERENCES


Edited by: S. V. Gordon