Probing Gene Expression in Live Single *Escherichia coli* Cells – One Molecule at a Time

Ji Yu, Jie Xiao*, Xiaojia Ren and X. Sunney Xie*[+]

Department of Chemistry and Chemical Biology, Harvard University,

Cambridge, MA 02138, USA

* These authors contributed equally.

+ Corresponding author, E-mail: xie@chemistry.harvard.edu

Abstract

We directly observe real-time production of single protein molecules in individual *E. coli* cells. A membrane-targeting sequence was fused to the gene of fast maturing yellow fluorescent protein (YFP) under the control of a *lac* promoter on the chromosome DNA. Gene expression under a repressed condition generates membrane-localized YFP molecules that can be detected with single molecule sensitivity. We report that protein production occurs in randomly distributed bursts and that the number of protein molecules in each burst follows a geometric distribution, consistent with a previous theoretical model. This study leads to quantitative understanding of the working of gene expression in live cells.

The central dogma of molecular biology states that DNA is transcribed into mRNA, which is then translated into protein. Ever since the pioneering work on the *lac* operon (1), genetic and biochemical studies (2-4) has conducted primarily with large populations of cells and molecules. Recently, many *in vitro* single-molecule experiments have allowed the probing of real-time dynamics and yield valuable mechanistic insights into macromolecules (5-8), including transcriptional (9) and translational machineries (10). In order to understand the working of these machineries in their physiological context, an ultimate quest is to probe gene expression on a single-molecule basis in a live cell. Single mRNA detection has been realized (11, 12). Here we demonstrate real-time monitoring of protein being generated one molecule at a time in live *E. coli* cells

Gene expression is by definition a single molecule problem, e.g. many genes exist as single copies in *E. coli*, being turned on and off to control many biological phenomena. Consequently, gene expression events are stochastic in nature (13-15). This stochasticity is manifested in both the distribution of the expression levels in a population of cells, and in the temporal fluctuation of an individual cell. The former has been the focus of several recent studies (16-18), while the latter has been difficult to probe to date due to technical barriers, but contains crucial dynamic information

The stochastic fluctuation is especially prominent at low levels of gene expression. Many genes produce only small numbers of proteins, but play essential roles in regulating biological functions. Current technologies are not sensitive enough to detect expression of these genes. In *E. coli*, gene expression is often controlled by repressors tightly bound to promoter regions of DNA. Here we seek to probe gene expression controlled by the *lac* promoter that is highly repressed (i.e., in the absence of an inducer

and in the presence of glucose), and use sensitive fluorescence reporters to follow protein production and promoter activity.

The most popular reporters for monitoring gene expression in live cells are green fluorescent protein (GFP) and its derivatives such as YFP (19-21). The long maturation time of GFP often limits the temporal resolution for monitoring gene expression. Therefore, we use an YFP variant, Venus, as the reporter because of its fast folding kinetics and short maturation time (22).

It is difficult to image a single GFP molecule in cytoplasm with a fluorescence microscope. The weak GFP signal spreads to the entire cytoplasm during the image acquisition time due to the molecule's fast diffusion, and is overwhelmed by cellular autofluorescence. On the other hand single GFP fusion protein molecules on cell membranes can be detected (23, 24) because their diffusion is slowed down. Therefore, we designed a fusion protein consisting of Venus and a membrane protein, Tsr, as the reporter for monitoring *lac* promoter activity. A well-studied methylation-dependent chemotaxis receptor protein (MCP) (25), Tsr, contains two transmembrane domains at its N terminus and a cytoplasmic domain at its C terminus, to which Venus is fused.

We constructed an *E. coli* strain SX4 in which a single copy of chimeric gene *tsr-venus* is incorporated into the *E. coli* chromosome, replacing the native *lac*Z gene. The endogenous *tsr* gene of *E. coli* is left intact. Since the endogenous *tsr* gene is highly expressed (25), a small amount of exogenous Tsr-Venus poses minimal perturbation to cells' normal functions. Using quantitative Western blotting assay (Fig. S2), we compared the expression levels of fully induced Tsr-Venus and Venus, both under the control of the *lac* promoter. No significant difference between the two strains was

observed, indicating that little perturbation to the expression of *venus* is introduced by the *tsr* sequence.

We first show the ability to detect single Tsr-Venus fluorescent protein molecules expressed in SX4 strain of *E. coli* cells (Fig. 1). Fig. 1A shows two diffraction-limited fluorescent spots (FWHM ~300 nm) in the left cell. A line cross-section of the fluorescence image along the cells' long axes shows the signal distinctly above the cells' autofluorescence background (Fig. 1C). We attribute each fluorescent peak to an individual Tsr-Venus molecule based on the abrupt disappearance of the signal because of photobleaching, which is characteristic of single molecules. Fig. 1D shows such a photobleaching time trace. Had the signal arises from multiple molecules, its disappearance would be in multiple steps. In addition, the fluorescence intensity of each peak is consistent with *in vitro* measurements of purified single Venus molecules (Fig. S1).

Fig. 2 is the sketch of the experimental design for live cell observations. Upon an infrequent and spontaneous dissociation event of the repressor from the operator region of DNA, transcription by RNA polymerase is initiated, generating one mRNA molecule. A few ribosome molecules bind to the mRNA, producing a burst of fusion protein molecules. These molecules can be detected after the completion of their assembly process, including protein folding, incorporation onto the inner cell membrane, and maturation of the Venus fluorophores. Meanwhile, the repressor quickly rebinds to the operator under the highly repressing conditions, and remains bound until the next event of spontaneous dissociation triggers another burst of protein production.

To count accurately the fusion protein molecules as they are continuously generated, we photobleach the Venus fluorophores after their detection. Specifically, we apply a 1200-ms laser exposure every 3 min. The laser power used is 0.3 kw/cm^2 with a photobleaching time constant of ~250 ms. Images are collected only in the first 100 ms, during which photobleaching is minimal, and are discarded in the following 1100 ms, because photobleaching causes stochastic variation in the integrated signal and decreases the signal to background ratio. The 3-min dwell time, which defines the temporal resolution, is chosen to avoid photo-damage to the cells. As a control, cells grown under such laser illumination in a temperature-controlled chamber on a microscope stage have a cell cycle time of ~55 min, the same as that in vigorously shaking M9 liquid culture without laser illumination.

The activity of the *lac* promoter in SX4 strain under the highly repressing condition was monitored for cells immobilized by an agarose gel pad of M9 media through several cell cycles. An epi-fluorescence microscope and CCD camera were used to image Venus molecules with 514 nm laser excitation, while differential interference contrast (DIC) images were taken simultaneously to record the cell contours during growth. Time-lapse movies were recorded (two shown in Movie S1 & S2). A sequence of images from one of them is shown in Fig. 3. In each fluorescence image, the fluorescent spots correspond to newly synthesized fluorescent molecules during the last 3 min. It is interesting to note that although Tsr is known to cluster at the cell pole, we observed many molecules around the center of the cell, indicating that these are newly synthesized protein molecules that haven't been localized to cell pole yet. Time traces of the fluorescent protein molecules along cell lineages are extracted from the time-lapse

fluorescence/DIC movies. Three time traces along particular cell lineages are shown in the bottom panels of Fig. 3. More than 60 time traces have been collected from repeated time-lapse microscopy measurements for statistical analyses.

Several interesting qualitative features are evident from these time traces. First, protein molecules are generated in clusters, which we define as gene expression bursts. Second, the number of protein molecules in each burst varies. Third, the bursts exhibit particular temporal widths. Three questions arise: Do these gene expression bursts occur randomly in time? What is the distribution of the number of protein molecules in each burst? And what is the origin of the temporal spread of the individual bursts? These questions are pertinent to the current discussion in the literatures in the context of stochasticity of gene expression (14-18). However, no direct measurements in real-time have been performed to date due to the lack of a sensitive live-cell assay.

To address the first question, the distribution of the number of gene expression bursts in one cell cycle for all cells is shown in Fig. 4A. The histogram is well fit with a Poisson distribution, which suggests that gene expression bursts are random and uncorrelated in time. We note that a weak cell cycle dependence of the burst frequency was observed (Fig. S4), but this should not change the Poissonian distribution of the gene expression bursts. The average number of bursts is 1.2 ± 0.3 per cell cycle, yielding an average time between two adjacent bursts of ~46±12 min. This time is compared with *in vitro* dissociation times of *lac* repressor from *lac* operator O_1(20-50 min)(26, 27), keeping in mind that the dissociation time in live cell might be different, and that each repressor dissociation event may not lead to successful transcription due to either temporary unavailability of RNA polymerase or failed transcription.

We tentatively attribute each gene expression burst to one copy of mRNA based on the fact that the transcription initiation time (28) (~10 seconds) is comparable or longer than the time during which the repressor is off the operator (a few seconds (26, 27)). However, we cannot experimentally rule out the possibility that multiple mRNAs were generated in the same burst.

To address the second question regarding the distribution of protein molecules per gene expression burst, Fig. 4B shows the histogram of the number of molecules per burst. We attribute the variation of the number to the finite cellular lifetime of an mRNA molecule. The degradation of mRNA is initiated by RNase E, which competes with the ribosome in binding to mRNA (29). Accordingly, it was proposed that the probability of generating n protein molecules from one mRNA follows a geometric distribution (14):

$$P(n) = p^n \cdot (1-p) \qquad (1)$$

where p is the probability of the ribosome binding and 1-p is that of RNase E binding to overlapping sites on mRNA (29). We note in deriving Eq. 1, it was assumed that a single copy of mRNA produces one gene expression burst. Had the number of mRNA molecules per burst followed a geometric distribution, the numbers of protein molecules per burst would still follow the geometric distribution (Supporting Text).

The histogram in Fig. 4B fits well with Eq. 1, yielding a value of 0.8 ± 0.1 for p with the mean of the distribution being 4.2 ± 0.5 molecules per gene expression burst. This number multiplied by the number of gene expression bursts per cell cycle (1.2) results in an average of 5.0 ± 0.8 molecules per cell. We experimentally confirmed the population averaged expression level to be 4.1 ± 1.8 molecules per cell by counting the molecules in ~300 individual cells under the microscope in a short data acquisition time

(100 ms). This result is also consistent with the measurement by quantitative Western Blotting assay (3.9 ± 2.3).

Our measured geometric distribution provides strong and first-of-a-kind experimental evidence for the simple model (14) of protein production and mRNA degradation described above. The geometric distribution of the protein numbers generated from each mRNA suggests that the survival times of the mRNA before RNase E binding follows a single exponential distribution. Although mRNA degradation is a complex process involving multiple enzymes, the single exponentially distributed survival probability indicates that there is only one rate-limiting step for inhibition of protein synthesis. To verify this, we made ensemble averaged measurements of the degradation of *tsr-venus* mRNA, which showed a single exponential decay with a time constant of 1.5 ± 0.2 min (Fig. S3), consistent with previous reports on many other genes in *E. coli*.

To address the third question about the spread of the gene expression bursts, we show in Fig. 4C the autocorrelation function of fluctuation in protein expression, $C^{(2)}(t)$, averaged from thirty different cell lineages from fifteen different movies. The single exponential fit of $C^{(2)}(t)$ gives a decay time constant of 7.0 ± 2.5 min, corresponding to the average width of the gene expression bursts. The entire protein production process consists of transcription, translation, and post-translational assembly that includes folding, membrane incorporation, and fluorophore maturation. Assuming one mRNA molecule per gene expression burst, we consider the following three facts in assigning the rate-limiting step for protein production. First, in *E. coli*, transcription and translation are coupled and the translation rate is ~15 amino acids per second. Therefore, the synthesis

of an 815-amino-acid fusion polypeptide is completed within 1 min. Second, the separation of two adjacent ribosomes on the mRNA is ~30 nm according to electron micrograph studies (30), implying that the separation between two consecutive polypeptides can be as short as ~3 s. Third, the *tsr-venus* mRNA lifetime is only ~1.5 min. Therefore we assign the 7-min burst width to the post-translational assembly.

We proved (Supporting Text) that under the condition that there is one rate-limiting step for the post translation assembly of the fusion protein,

$$C^{(2)}(\tau) = (\frac{sp}{1-p})^2[1+\frac{\kappa}{s}\exp(-\kappa\tau)] \qquad (2)$$

where s is the average frequency of the gene expression burst, κ is the rate constant of Tsr-Venus assembly, and p is the probability of ribosome binding as defined before. The fitting of Eq. 2 (Fig. 4C) gives $1/\kappa$ ~7.0 ± 2.5 min, which is associated with the rate limiting step among the folding, membrane incorporation, and fluorophore maturation steps. The fitting to Eq. 2 also results in s of 29 ± 8 min^{-1}, consistent with the average number of gene expression burst per cell cycle of 1.2 ± 0.3 (Fig. 4A) within experimental error, as well as p of 0.7 ± 0.1, consistent with the value of 0.8 ± 0.1 determined from Fig. 4B. These results provide examples of extracting cellular kinetic information from stochastic trajectories. We tentatively assign the rate constant, κ, to the fluorophore maturation process, though we cannot experimentally rule out the other two possibilities. In principle, a faster maturing construct would allow the kinetic measurements of protein folding and membrane incorporation.

Gene expression, central to life processes, is intrinsically a single molecule process. Yet, it has not been studied on a single molecule basis in individual cells due to technical barriers. Our single-molecule real-time assay in live cells opens up many new

possibilities for probing biochemical reactions in concert with physiological environments. For example, similar experiments with reporters of different colors would allow studies of the interplay of multiple genes. In eukaryotes, gene expression and regulation are more complex, and the single-molecule approach may allow the unraveling of complicated kinetics and mechanisms. Furthermore, the single molecule sensitivity of the assay will allow gene expression profiling of low copy number proteins or repressed promoters currently not accessible by genomics and proteomics technologies such as microarray (31) and mass spectrometry (32). We anticipate similar single-molecule assays in living cells will yield new knowledge of not only gene expression but also other fundamental biological processes.

References:

1. F. Jacob, J. Monod, *Cold Spring Harbor Symposia on Qantitative Biology* **26**, 193-211 (1961).

2. J. Beckwith, *Science* **156**, 597-604 (1967).

3. W. Gilbert, B. Muller-Hill, *Proc Natl Acad Sci USA* **58**, 2415-21. (1967).

4. M. Ptashne, A. Gann, *Genes & Signals* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2002).

5. X. S. Xie, H. P. Lu, *J. Biol. Chem.* **274**, 15967-70. (1999).

6. C. Bustamante, Z. Bryant, S. Smith, *Nature* **421**, 423-7 (2003).

7. A. Ishijima, T. Yanagida, *Trends in Biochemical Sciences* **26**, 438-444 (2001).

8. S. Weiss, *Science* **283**, 1676-83. (1999).

9. M. Wang, et al., *Science* **282**, 902-7 (1998).

10. S. Blanchard, R. Gonzalez, H. Kim, S. Chu, J. Puglisi, *Nat .Struct. Mol. Biol.* **11**, 1008-14 (2004).

11. Y. Shav-Tal, et al., *Science* **304**, 1797-800 (2004).

12. I. Golding, E. Cox, *PNAS* **101**, 11310-5 (2004).

13. O. Berg, *J. Theor. Biol.* **71**, 587-603 (1978).

14. H. H. McAdams, A. Arkin, *Proc Natl Acad Sci USA* **94**, 814-9. (1997).

15. J. Paulsson, *Nature* **427**, 415-8 (2004).

16. N. Barkai, S. Leibler, *Nature* **403**, 267-8 (2000).

17. E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, A. van Oudenaarden, *Nature Genetics* **31**, 69-73. (2002).

18. M. B. Elowitz, A. J. Levine, E. D. Siggia, P. S. Swain, *Science* **297**, 1183-6. (2002).

19. R. J. Bongaerts, I. Hautefort, J. M. Sidebotham, J. C. Hinton, *Methods Enzymology* **358**, 43-66. (2002).

20. R. Y. Tsien, *Annual Review of Biochemistry* **67**, 509-44. (1998).

21. M. Chalfie, Y. Tu, G. Euskirchen, W. W. Ward, D. C. Prasher, *Science* **263**, 802-5. (1994).

22. T. Nagai, et al., *Nature Biotechnology* **20**, 87-90. (2002).

23. P. H. Lommerse, et al., *Biophysical J.* **86**, 609-16. (2004).

24. J. Deich, E. M. Judd, H. H. McAdams, W. E. Moerner, *Proc Natl Acad Sci USA* **101**, 15921-6 (2004).

25. S. L. Mowbray, M. O. J. Sandgren, *Journal of Structural Biology* **124**, 257-275 (1998).

26. K. Bondeson, A. Frostellkarlsson, L. Fagerstam, G. Magnusson, *Analytical Biochemistry* **214**, 245-251 (1993).

27. W. Hsieh, P. Whitson, K. Matthews, R. Wells, *J. Biol. Chem.* **262**, 14583-91 (1987).

28. R. M. Saecker, et al., *J. Mol. Biol.* **319**, 649-671 (2002).

29. O. Yarchuk, N. Jacques, J. Guillerez, M. Dreyfus, *J. Mol. Biol.* **226**, 581-96. (1992).

30. O. L. Miller Jr, B. Hamkalo, C. A. Thomas, *Science* **169**, 392-5 (1970).

31. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467-70. (1995).

32. R. Aebersold, M. Mann, *Nature* **422**, 198-207. (2003).

33 We thank L. Xun, P. Choi, N. Friedman, L. Cai and R. Losick for helpful discussions and critical reading of the manuscript, and J. Hearn for technical assistance. This work was supported by National Institute of Health (NIH) Director's Pioneer Award program, a NIH R21 grant, and Department of Energy's Genome To Life program. JY acknowledges a Genome-related Research Award from Merck.

Figure Captions

Fig. 1. Single-molecule detection of a fluorescent fusion protein, Tsr-Venus, in live *E. coli* cells. (A) Fluorescence and (B) differential interference contrast images of two *E. coli* cells (strain SX4) expressing Tsr-Venus. Two single fusion protein molecules were detected as diffraction-limited fluorescent spots (full width at half maximum at ~300 nm) in the left cell. The fluorescence image is taken with 514-nm laser excitation and a 100-ms exposure time at 0.3 kW/cm^2. (C) Line cross-section of the fluorescence signal along long axes of the two *E. coli* cells. (D) Fluorescence time trace of a single Tsr-Venus molecule in an *E. coli.* Cell, showing abrupt photobleaching (40 ms exposure at 0.5 kW/cm^2).

Fig. 2. Experimental design for live-cell observations of gene expression. A fusion protein Tsr-Venus is expressed under the control of *lac* repressor, which binds tightly to the *lac* operator on DNA. Transcription of one mRNA by an RNA polymerase results from an infrequent and transient dissociation event of repressor from DNA. Multiple copies of protein molecules are translated from the mRNA by ribosomes. Upon being assembled into *E. coli*'s inner membrane, Tsr-Venus protein molecules can be detected individually in a fluorescence microscope.

Fig. 3. Real-time monitoring the expression of *tsr-venus* under the control of repressed *lac* promoter. (Top Panel) Sequence of fluorescent images (yellow) overlaid with simultaneous DIC images (gray) of *E. coli* cells expressing Tsr-Venus on agarose gel pad

of M9 medium. The cell cycle is 55 ± 10 min in a temperature-controlled chamber on a microscope stage. The eight frames are from the time-lapse fluorescence movie taken over 195 min with 100-ms laser exposures (0.3 kW/cm^2) every 3 min. An 1100-ms exposure is applied following each image collection to photobleach the Venus fluorophores. See supporting online material for the full movie. (Bottom Panel) Time traces of the expression of Tsr-Venus protein molecules (left) along three particular cell lineages (right) extracted from the time-lapse fluorescence/DIC movie of top panel. The time resolution is 3 min. The vertical axis is the number of protein molecules newly synthesized during the last three minutes. The dotted lines mark the cell division times. The time traces show that protein production occurs in random bursts, within which variable numbers of protein molecules are generated. Each gene expression burst has a finite spread of ~3-15 min.

Fig. 4. Statistical analyses of the protein production time traces. (A) Histogram (gray bars) of the number of expression events per cell cycle. The data is fit well to a Poisson distribution (solid line) with the average number of gene expression burst per cell cycle being 1.2 ± 0.3. (B) Distribution of the number of fluorescent protein molecules detected in each gene expression burst, which follows geometry distribution (solid line), giving the probability of ribosome binding of 0.81 ± 0.05 and the average number of molecules per burst of 4.2. (C) Autocorrelation function of the protein production time traces calculated according to Eq. S9. The result is averaged from 30 individual cell lineages because of the insufficient statistics of a single time trace. The fitting to Eq. 2 (solid line) gives a

time constant of 7.0 ± 2.5 min, which is attributed to posttranslational assembly of the fluorescent fusion protein.

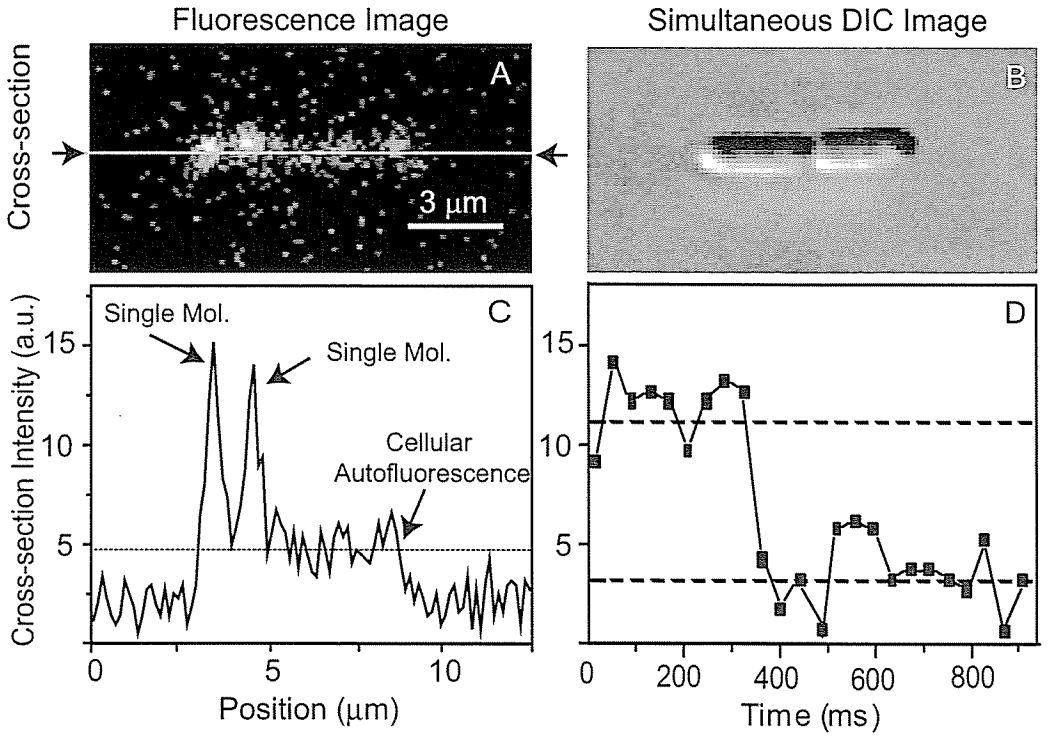
Fluorescence Image
Simultaneous DIC Image
Cross-section
A
B
3 μm
C
D
Single Mol.
Single Mol.
Cellular
Autofluorescence
Cross-section Intensity (a.u.)
15
10
5
0
5
10
Position (μm)
200
400
600
800
Time (ms)

Fig. 1

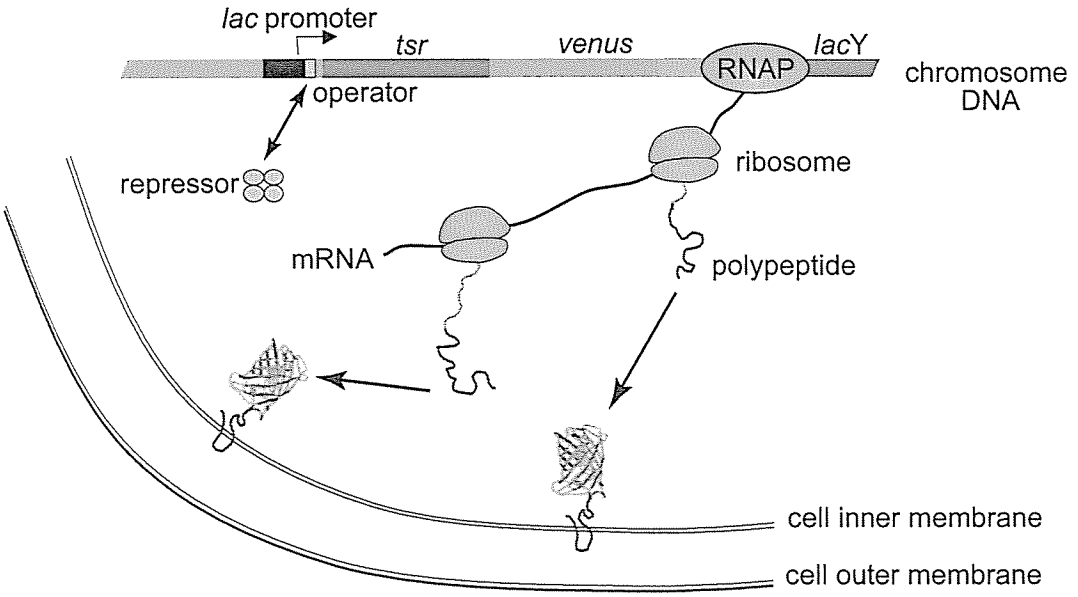
lac promoter
tsr
venus
lacY
RNAP
chromosome DNA
operator
repressor
ribosome
mRNA
polypeptide
cell inner membrane
cell outer membrane

Fig. 2

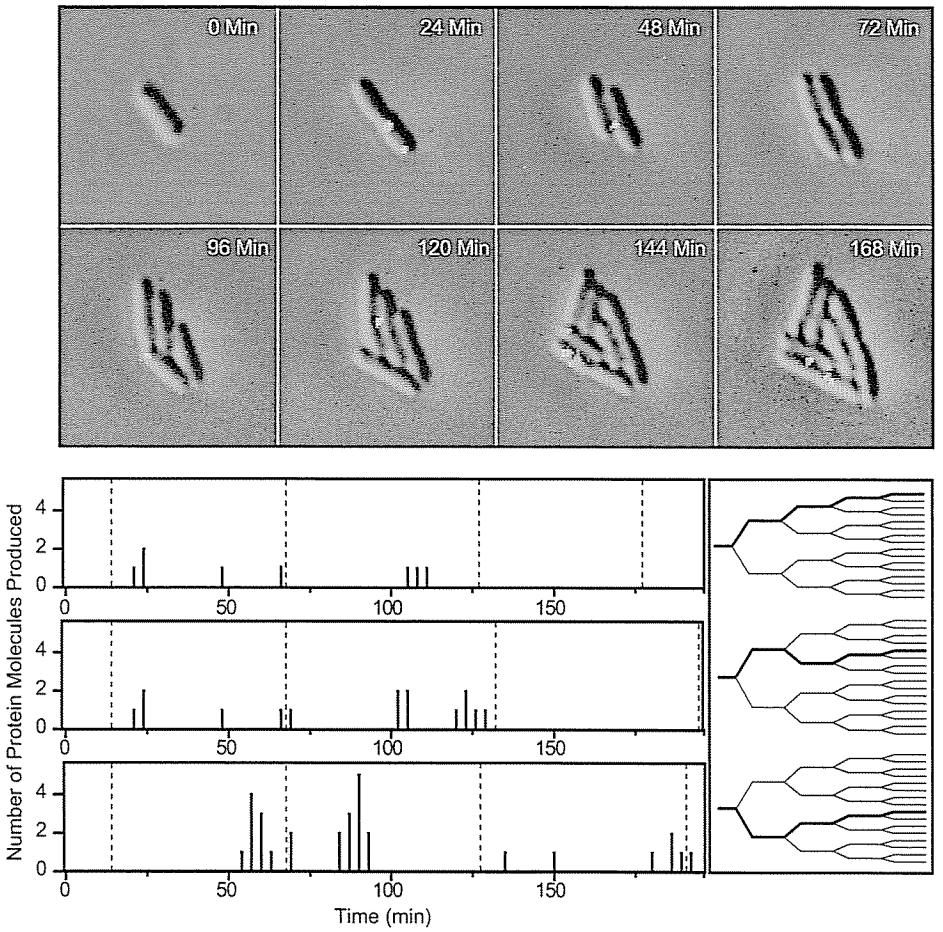
0 Min
24 Min
48 Min
72 Min
96 Min
120 Min
144 Min
168 Min
Number of Protein Molecules Produced
0
2
4
0
50
100
150
Time (min)

Fig. 3

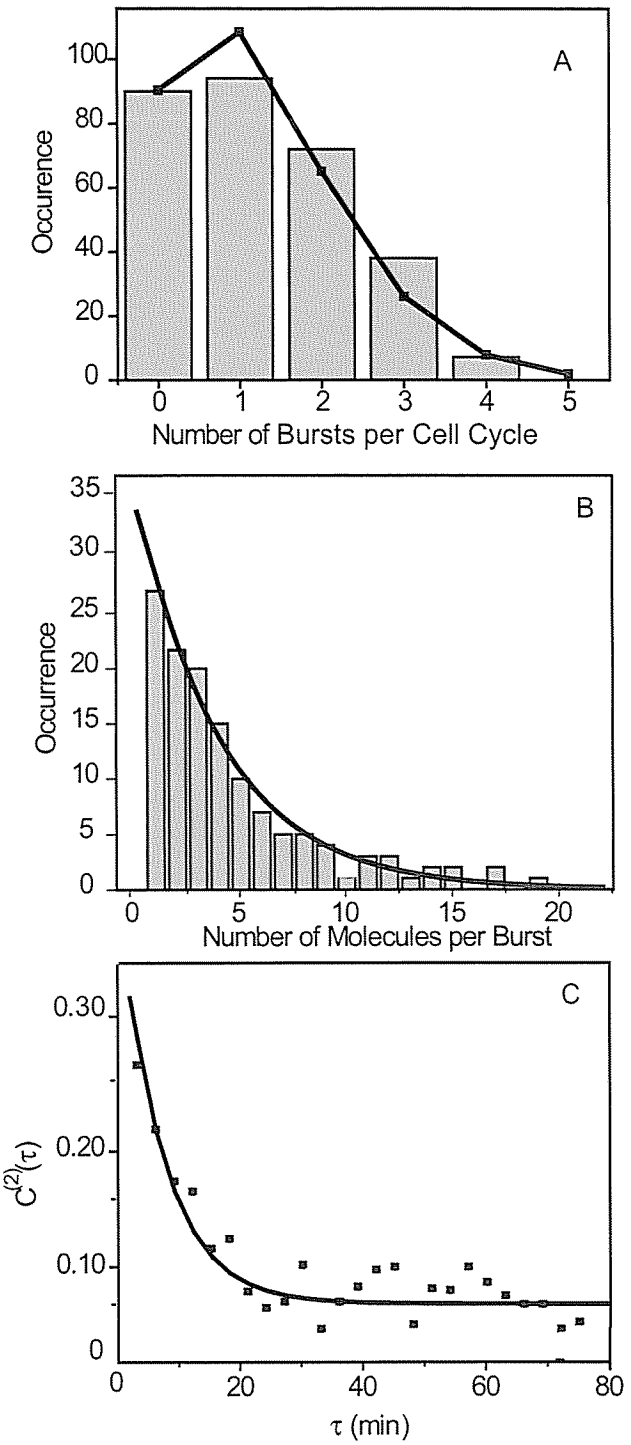
A
Occurence
100
80
60
40
20
0
0
1
2
3
4
5
Number of Bursts per Cell Cycle
B
Occurrence
35
30
25
20
15
10
5
0
0
5
10
15
20
Number of Molecules per Burst
C
$C^{(2)}(\tau)$
0.30
0.20
0.10
0
0
20
40
60
80
τ (min)

Fig. 4