ORIGINAL PAPER

EBSA

Biophysics in Europe

The statistics of protein expression ratios for cellular fluorescence studies

Elizabeth M. Smith · Joachim D. Mueller

Abstract Fluorescence studies of cellular protein-protein interactions commonly employ transient cotransfection to express two proteins carrying distinct fluorescent labels. Because transiently transfected cells differ significantly in their expression level, the concentration ratio of the two expressed proteins varies, which in turn influences the measured fluorescence signal. Knowledge of the statistics of protein expression ratios is of considerable interest both from a fundamental point of view and for cellular fluorescence studies. Despite the perceived randomness of transient transfection, we were able to develop a quantitative model that describes the average and distribution of the protein expression ratio from a cell population. We show that the expression ratio is proportional to the molar plasmid ratio and relate the distribution to the finite number of active plasmids in the cell. The process of cationic lipid-mediated transfection is explored in more detail. Specifically, the influence of lipoplexes on the statistics of the expression ratio is examined. We further demonstrate that the transfection model provides a quantitative description of fluorescence fluctuation experiments, where only a fraction of the proteins are labeled.

E. M. Smith · J. D. Mueller (✉)
School of Physics and Astronomy, University of Minnesota,
116 Church St., SE, Minneapolis, MN 55455, USA
e-mail: mueller@physics.umn.edu

J. D. Mueller
Institute for Molecular Virology, University of Minnesota,
Minneapolis, MN 55455, USA

J. D. Mueller
Department of Biomedical Engineering, University
of Minnesota, Minneapolis, MN 55455, USA

Introduction

Transient transfection is an important technique for the transfer of DNA into cultured mammalian cells. Because transient transfection is simple, fast, and permits the convenient transient expression of foreign genes, it has become an essential tool for cell biological studies. A powerful approach for the study of protein behavior inside a cell is the direct visualization of the protein by tagging it with a green fluorescent protein (GFP) or one of its derivatives (Day and Davidson 2009; Snapp 2009). The fluorescent signal from the tagged protein provides important information about localization, mobility, transport, and interactions of the protein (Ciruela 2008; Day and Schaufele 2008). Thus, transient transfection of GFP-tagged 'fusion' proteins is widely utilized in live cell studies.

Experiments that probe protein interactions, such as fluorescence resonance energy transfer (FRET) and fluorescence correlation spectroscopy (FCS), often rely on dual-color fluorescence studies, where two proteins tagged with distinct fluorescent proteins are coexpressed in the same cell (Berland 2004; Gambin and Deniz 2010; Haustein and Schwille 2007). While it is known that transient transfection results in a large cell-to-cell variability of protein expression, the properties of the expression ratio between two coexpressed proteins are relatively unknown. This ratio is important for cellular fluorescence studies because it affects the measured fluorescence signal and also provides insights into fundamental aspects of transient transfection.

Springer

In this paper, we aim to quantify transient cotransfection through the introduction of a simple model that describes the mean, standard deviation, and distribution of the protein coexpression ratio. We explicitly relate the mean of the expression ratio to the plasmid DNA (pDNA) molar ratio of the transfection solution and extend the model to include the variability of protein expression. The model is compared against experimental data, and the simplifying assumptions of the model are discussed. Our study also sheds light on the cationic lipid-mediated transfection process. Although this transfection method was introduced over 2 decades ago (Felgner et al. 1987), some details are still unclear, and only limited work has examined the process of quantitative cotransfection (Ma et al. 2007). Here, we specifically investigate the role of lipoplexes in the transfection process. We demonstrate that only very few lipoplexes contribute successfully to gene expression and discuss the resulting consequences for protein expression ratios. Although our model simplifies many details of the transfection process, it still reproduces the data remarkably well.

Modeling of the relation between the expression ratio and the molar plasmid ratio of the transfection solution is also extremely useful in application. With the model, it is now possible to adjust the expression ratio in a predictive fashion, which is desirable for optimizing protein interactions in fluorescence studies. Our model predicts that this adjustment is limited to a finite range, as is dictated by properties of the transient transfection process. Additionally, being able to statistically describe the cell-to-cell variability of the protein expression ratio allows us to study its influence on the outcome of specific cellular fluorescence studies.

We introduce and apply our model to a 'bright and dark' fluorescence fluctuation experiment, where only a fraction of the proteins carry a fluorescent label. The unlabeled protein is non-fluorescent and cannot be directly observed in fluorescence fluctuation spectroscopy (FFS) experiments. For this reason, we frequently refer to the unlabeled protein as a dark protein, and experiments that coexpress a labeled protein together with an unlabeled protein as bright and dark experiments. Because labels potentially interfere with the assembly of protein complexes, it is useful to establish control experiments that detect such an artifact. A bright and dark experiment has the potential to identify label artifacts, but requires knowledge of the statistics of protein expression ratios because the amount of dark protein directly influences the measured signal. We demonstrate that the transfection model establishes a framework for interpretation of FFS bright and dark experiments.

Materials and methods

Experimental setup

Experiments were performed using a Zeiss Axiovert 200 microscope (Zeiss, Gottingen, Germany) with a Plan-Apochromate $63\times$ oil immersion objective (N.A. = 1.4) to focus the laser beam and collect the fluorescent emission. A Ti:sapphire laser (Tsumami, Spectra-Physics, Mountain View, CA), pumped by an intracavity doubled Nd:YVO$_4$ laser (Millennia vs. Spectra-Physics), served as the two-photon excitation source. All measurements were taken with a 1,000-nm excitation wavelength and an average power after the objective of 0.25 mW. For dual-channel measurements, a dichroic mirror with a center wavelength of 525 nm split the fluorescence emission into two detection channels. The fluorescence in each channel was detected by an avalanche photodiode (APD, SPCM-AQ-141, Perkin-Elmer, Dumberry, Quebec). The dichroic mirror was removed during single-channel measurements. The TTL output from each APD was recorded by a PCI data acquisition card (ISS, Champaign, IL). Data were acquired at a frequency of 20 kHz for ~ 10 s for intensity fraction measurements and for ~ 1 min for brightness measurements. The photon counts were analyzed with programs written in IDL 7.1 (Research Systems, Boulder, CO).

Sample preparation

pEGFP-C1 and pEYFP-C1 plasmids were purchased from Clontech (Clontech, Palo Alto, CA). Tandem dimeric EGFP (EGFP$_2$), RXRLBD-EYFP, RXRLBD-EGFP, and RARLBD-EYFP plasmids were constructed as previously described (Chen et al. 2003, 2005). EGFP-endophilin A2 (Endo-EGFP) plasmids were a gift from Dr. Joseph Albanesi (University of Texas Southwestern Medical Center). All sequences were verified by automatic sequencing.

CV-1 cells were obtained from ATCC (Manassas, VA) and maintained in 10% fetal bovine serum (Hycolone Laboratories, Logan, UT) and DMEM medium. Cells were subcultered in eight-well coverglass chamber slides (Naglenunc International, Rochester, NY) 12 h before transfection. Transient transfections were carried out using TransFectin (BioRad, Hercules, CA) according the manufacturer's instructions 24 h prior to measurement. For cotransfections, a standard protocol was followed with the two plasmid types mixed together at a given mole ratio prior to adding TransFectin. In one specific case, a second cotransfection protocol was employed where TransFectin was added to specific plasmids (rather than the plasmid mixture) to form pDNA/lipid complexes that only contained one type of pDNA. The pDNA/lipid complexes

containing each type of pDNA were then mixed together at a given mole ratio. In each transfection protocol, ~ 0.2 μg of total pDNA was added to each well. Immediately before measurement, the growth medium was replaced with Dulbecco's phosphate-buffered saline (PBS) with calcium and magnesium (Biowhittaker, Walkerville, MD). Measurements of EGFP, EYFP, EGFP$_2$, RXRLBD-EYFP, RXRLBD-EGFP, and RARLBD-EYFP were carried out in the cell nucleus, while measurements of Endo-EGFP and Endo were carried out in the cell cytoplasm.

Theory

Dual color protein expression fraction

The average detected fluorescence intensity $\langle F \rangle$ is given by

$$\langle F \rangle = \lambda N \tag{1}$$

where the brightness λ is the average photon count rate for a single molecule, and N is the detected number of molecules (Müller 2004). In two channel measurements, the emitted fluorescence from a single fluorophore is split by a dichroic filter into two detection channels ("red" and "green"). The average fluorescence intensity in the green channel is $\langle F^{(g)} \rangle = \lambda^{(g)} N$ and that of the red channel is $\langle F^{(r)} \rangle = \lambda^{(r)} N$, where $\lambda^{(g)}$ and $\lambda^{(r)}$ are the brightness values of the green and red channel, respectively. Because both detectors see the same number of molecules, the fluorescence intensity fraction of the red channel,

$$f_I = \frac{F^{(r)}}{F^{(r)} + F^{(g)}} = \frac{\lambda^{(r)}}{\lambda^{(r)} + \lambda^{(g)}} \tag{2}$$

equals the brightness fraction. The emitted color of the fluorescent protein together with the dichroic filter determines the intensity fraction.

In a mixture of two non-interacting proteins with distinct labels, the intensity fraction of the red channel is now determined by Chen et al. (2005)

$$f_I = \frac{F^{(r)}}{F^{(r)} + F^{(g)}} = \frac{\lambda_G^{(r)} N_G + \lambda_Y^{(r)} N_Y}{\lambda_G^{(g)} N_G + \lambda_Y^{(g)} N_Y + \lambda_G^{(r)} N_G + \lambda_Y^{(r)} N_Y} \tag{3}$$

where the subscripts G and Y refer to the fluorescent proteins, EGFP and EYFP. For this study we assume no FRET is occurring between the two proteins. Chen et al. (2005) describes how, if necessary, FRET is taken into account. By rewriting Eq. 3 we relate the protein expression fraction to experimentally determined parameters,

$$f_N = \frac{N_Y}{N_Y + N_G} = \frac{f_I(\lambda_G^{(g)} + \lambda_G^{(r)}) - \lambda_G^{(r)}}{f_I(\lambda_G^{(g)} + \lambda_G^{(r)} - \lambda_Y^{(g)} - \lambda_Y^{(r)}) - \lambda_G^{(r)} + \lambda_Y^{(r)}}. \tag{4}$$

The calibration parameters $(\lambda_G^{(r)}, \lambda_G^{(g)}, \lambda_Y^{(r)}, \lambda_Y^{(g)})$ are the brightness values of the fluorescent proteins (G = EGFP, Y = EYFP) in each detection channel (r = 'red' channel, g = 'green' channel) and are determined from separate FFS measurements of cells expressing EGFP or EYFP alone. For example, $\lambda_G^{(r)}$ and $\lambda_G^{(g)}$ are the brightness values for EGFP in the red and green channel, respectively. The brightness values for EGFP and EYFP in each channel account for the differences in detected photon count rates; when used in combination with the fluorescent intensity fraction (Eq. 3) the protein expression fraction for any cell can be determined.

Model for protein expression ratios

We introduce a simple linear model to describe the molar amount of protein v_P expressed in a cell,

$$v_P = x \cdot \eta \cdot N_A \tag{5}$$

where N_A presents the number of active plasmids in the cell, η is an efficiency factor, and x takes environmental and cell-specific factors that influence protein expression into account. The amount of expressed protein is expected to be directly proportional to the number of active plasmids N_A presuming that sufficient cellular resources are present. The expression efficiency of individual genes within a plasmid varies. For example, mRNA translational efficiency is influenced by its sequence (Voges et al. 2004). These differences are described by the efficiency factor η. More generally, the factor η can be used to account for transcription and translation-related factors influencing expression efficiency. For example, we expect that changing the promoter affects the value of the efficiency factor. However, this paper only considers plasmids with the same promoter. Protein expression depends on many additional environmental and cell-specific factors. For example, the entry of plasmid DNA into the nucleus after transfection is expected to vary from cell to cell, which would lead to a distributed start time of protein production in a cell population. The compound effect of these diverse factors on the amount of expressed protein is symbolically described by the parameter x.

The presence of two types of plasmids (pDNA$_1$ and pDNA$_2$) in a cell leads to the coexpression of two proteins. The ratio of the molar amount of expressed protein $r_P = v_1/v_2$, is according to Eq. 5 directly proportional to the active plasmid number ratio $r_A = N_{A1}/N_{A2}$,

$$r_P = \frac{x_1}{x_2} y \cdot r_A \approx y \cdot r_A \tag{6}$$

with $y = \eta_1/\eta_2$. It is expected that $x_1 \approx x_2$ in the same cell. This statement assumes that the variance in the ratio of x_1 and x_2 is sufficiently small such that the protein ratio r_P is

Table 1 Definition of parameters

Symbol	Definition
r_P	Protein expression ratio
r_S	Molar ratio of pDNA in solution
r_A	"Active" plasmid ratio
η_i	Factor accounting for protein specific expression efficiency of $pDNA_i$
y	Efficiency ratio = η_1/η_2
f_P	Protein expression fraction
f_S	Molar fraction of pDNA in solution
f_A	"Active" plasmid fraction
b	Normalized brightness

mainly determined by y and r_A. A summary of the relevant parameters can be found in Table 1. The relative protein ratio in terms of expressed protein concentration $r_P = c_1/c_2$, is experimentally measurable. In contrast, neither y nor r_A is directly measurable. The only other experimental quantity under our control is the molar mixing ratio r_S of pDNA in solution. Transfection leads to the uptake of plasmid by the cell. Only a small fraction of the transfected plasmids will ultimately be involved in active gene expression (Hama et al. 2006, 2007). We refer to such plasmids as active. The events that lead to an active plasmid are not yet fully understood, as will be discussed at a later point. Because very few plasmids become active, small number fluctuations become important. Thus, the ratio r_A of active plasmids of $pDNA_1$ and $pDNA_2$ will vary across cells. However, while the ratio r_A varies, the average of the active plasmid ratio equals the molar mixing ratio of pDNA in solution $\langle r_A \rangle = r_S$. This statement reflects that cells treat the two types of plasmids as equal. Neither uptake nor activation of a plasmid is expected to depend on the particular type of plasmid ($pDNA_1$ or $pDNA_2$), as long as the plasmids are of comparable length. Thus, the average active plasmid ratio must reflect the plasmid ratio in solution. This statement together with Eq. 6 predicts that the average protein ratio from a population of cells will be proportional to the molar ratio of pDNA in solution scaled by the relative efficiency factor.

$$\langle r_P \rangle \approx y \cdot \langle r_A \rangle = y \cdot r_S \tag{7}$$

This linear relationship can be experimentally tested, because both r_P and r_S can be measured.

Equation 7 describes the average behavior. As mentioned earlier the active plasmid ratio will vary from cell to cell. Thus, according to Eq. 6 the expressed protein ratio r_P will reflect these variations in r_A. The probability distribution function (pdf) of both quantities is related by

$$\text{pdf}_P(r_P)dr_p = \text{pdf}_A(r_A)dr_A \tag{8}$$

The probability distribution function $\text{pdf}_P(r_P)$ of the expressed protein ratio can be measured, but we require a model for the probability distribution function $\text{pdf}_A(r_A)$ of the active plasmids.

It is advantageous to use fractions instead of ratios for graphing and comparing distributions. Thus, we convert each ratio r into its corresponding fraction $f = r/(1 + r)$. By transforming Eq. 7 into fractions we get

$$\langle f_P \rangle \approx \frac{y\langle f_A \rangle}{1 + \langle f_A \rangle(y - 1)} = \frac{y \cdot f_S}{1 + f_S(y - 1)} \tag{9}$$

for the average protein fraction $f_P = r_P/(1 + r_P)$. The equivalent transformation for the probability distribution functions of Eq. 8 leads to

$$\text{pdf}_P(f_P)df_P = \text{pdf}_A(f_A)df_A \tag{10}$$

for the pdf of the protein expression fraction. Note that the distributions of f_P and f_A become identical for the special case of $y = 1$,

$$\text{pdf}_P(f_P) = \text{pdf}_A(f_A) \tag{11}$$

Results and discussion

Average protein expression fraction

We first examine the relationship between the plasmid mixing fraction f_S and the average protein expression fraction $\langle f_P \rangle$. These experiments provide a first test of the simple linear model postulated in Eq. 7. The initial experiments are performed on the fluorescent proteins EGFP and EYFP as they will serve as the fluorescent markers of all subsequent fusion proteins included in this study. The plasmids pEGFP and pEYFP are mixed at specific mole fractions f_S. Each plasmid mixture is subsequently used in the transfection of CV-1 cells. For each population of cells transfected with a specific plasmid mixing fraction, the fluorescence intensity of more than 20 cells was recorded in a dual-color setup. The protein expression fraction, $f_p = N_{EYFP}/(N_{EYFP} + N_{EGFP})$, was determined individually for each measured cell. Figure 1a shows the average protein expression fraction $\langle f_P \rangle$ versus the plasmid mixing fraction f_S. The data suggest a linear relationship between both fractions. A fit of the data to our proposed model (Eq. 9) reproduces the data with an efficiency ratio of $y = 1.04$. An efficiency ratio close to one is expected because the two plasmids contain the same promoter region, and the genetic sequence and structure of each protein is nearly identical.

We further test our model using fusion proteins. For the initial experiments we chose the ligand-binding domain of the retinoid X receptor protein (RXRLBD) and label it with either EGFP or EYFP. The plasmids pEGFP-RXRLBD and
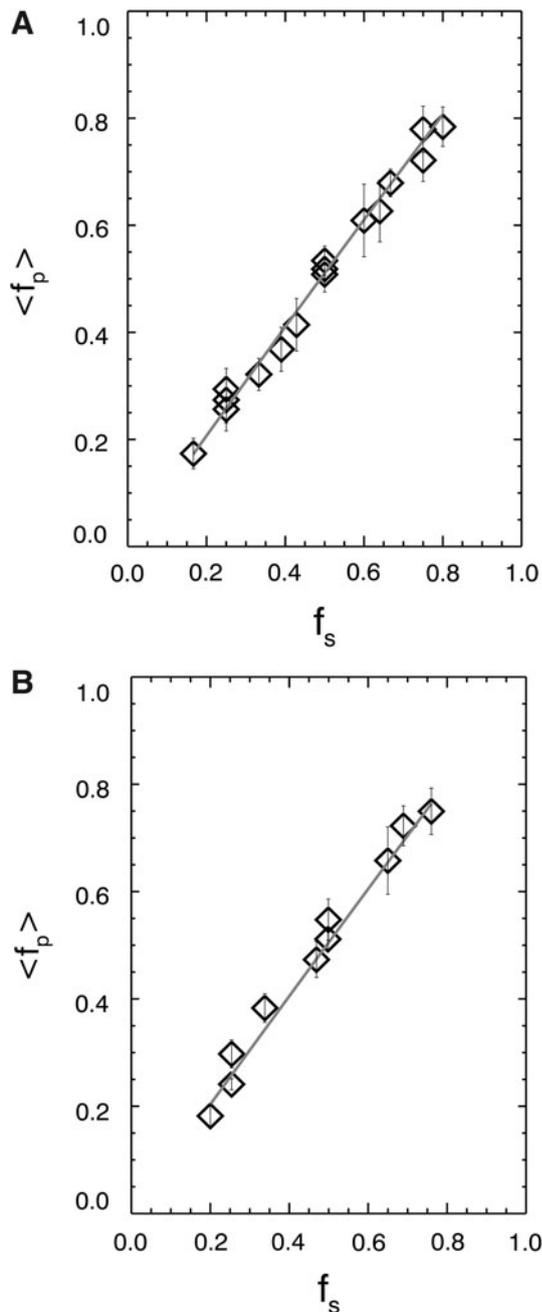
Fig. 1 Average protein expression fraction $\langle f_P \rangle = N_{P1}/(N_{P1} + N_{P2})$ as a function of the plasmid mole fraction $f_S = v_{p1}/(v_{p1} + v_{p2})$. Each data point represents the average of the expression fraction from 20 to 40 cells. **a** Data from cells expressing EYFP and EGFP with $P1 = EYFP$, $P2 = EGFP$, $p1 = pEYFP$, and $p2 = pEGFP$. The fit recovered an efficiency ratio of 1.08. **b** Data from cells expressing RXRLBD-EYFP and RXRLBD-EGFP with $P1 = RXRLBD\text{-}EYFP$, $P2 = RXRLBD\text{-}EGFP$, $p1 = pRXRLBD\text{-}EYFP$, $p2 = pRXRLBD\text{-}EGFP$. The fit of the model recovered an efficiency ratio of 1.03

pEYFP-RXRLBD, which essentially have identical sequences, are mixed at specific mole fractions, and the average protein expression fractions are calculated from the intensity fractions of more than 20 cells. Figure 1b

shows the results for the average protein versus plasmid fraction. The data again suggest a linear relationship, and a fit to our model (Eq. 9) yields an efficiency ratio of 1.02, which confirms that both proteins have nearly identical expression efficiencies.

We now examine two proteins that differ substantially in their sequence. In this case we still expect the relationship between pDNA mixing and protein expression to be given by Eq. 9, but with an efficiency ratio that may be different from one. We experimentally test this hypothesis by expressing the proteins EGFP and RXRLBD-EYFP. As in the previous experiments, plasmids for each protein are mixed at specific mole fractions, and the average protein expression fraction is calculated from a sample of more than 20 cells. Figure 2a shows the average protein expression fraction versus plasmid mixing fraction for this EGFP and RXRLBD-EYFP protein pair. The expressed protein fraction varies nonlinearly with the plasmid fraction, which indicates an efficiency ratio of $y \neq 1$. A fit of the data by Eq. 9 determines $y = 0.22$. We repeated the experiment for a second protein pair by expressing EGFP and RARLBD-EYFP (retinoic acid receptor). Again, we observe a nonlinear relationship between the protein and plasmid fractions (Fig. 2b). Fitting the data to Eq. 9 yields $y = 0.19$.

The insets in Fig. 2 show protein and plasmid ratios instead of fractions to highlight the underlying linear relationship postulated in Eq. 7. The average protein ratio versus the plasmid mixing ratio for each protein pair is given together with a solid line representing Eq. 7 with the previously fit y values as the linear scaling factor. From the insets in Fig. 2, one can directly see that the slope (y) is less than 1 for the dissimilar protein pairs; consequently, the relationship between protein and plasmid fractions in the main plot is non-linear. The protein fraction plots of Figs. 1 and 2 are constrained by the fractional endpoints $f_s = f_p = 0$ and $f_s = f_p = 1$. Only $y = 1$ results in a linear graph with a slope of 1 (Fig. 1). For $y < 1$, the protein represented in the numerator is less easily expressed compared to the other protein such that large changes in the plasmid mixture are necessary to produce small changes in the expressed protein level. Consequently, this leads to an initial linear region of the data with a slope less than one followed by a region of increasing slope as we approach the endpoint $f_s = f_p = 1$, which results in a positive curvature of the data. In contrast, for $y > 1$, the protein represented in the numerator is more easily expressed compared to the other protein; this introduces a negative curvature in a plot of f_P versus f_S with a slope initially greater than one, which decreases as we approach the endpoint $f_s = f_p = 1$. A fit of the data to Eq. 9 determines y by taking these changes in slope and curvature into account.

Additionally, we investigated the influence of label selection on the relationship between the plasmid mixing
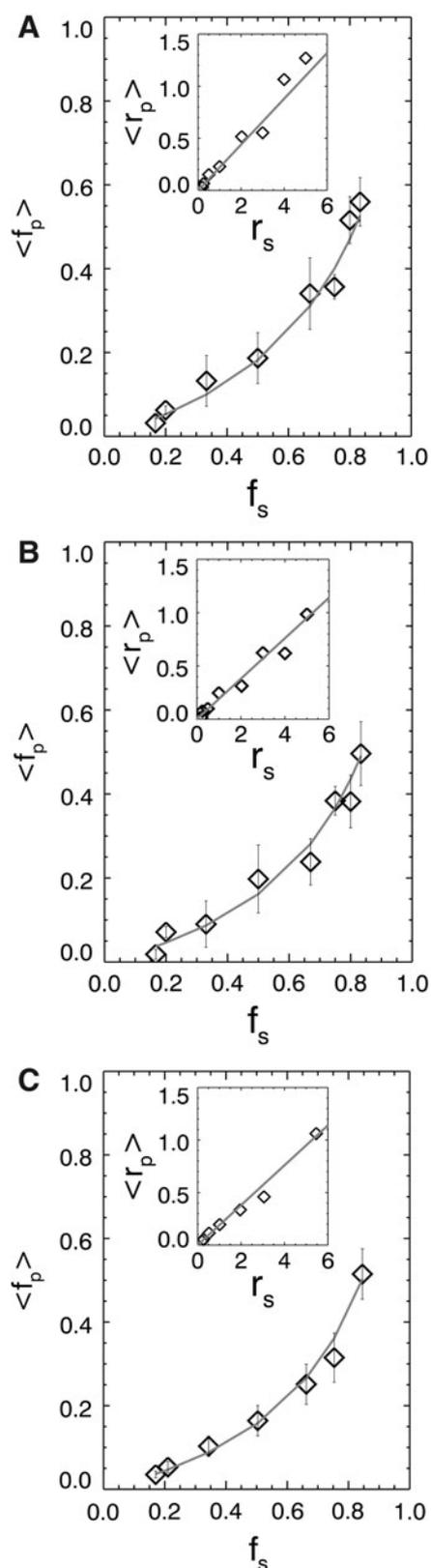
◀ **Fig. 2** Average protein expression fraction $\langle f_P \rangle = N_{P1}/(N_{P1} + N_{P2})$ as a function of the plasmid mole fraction $f_S = v_{p1}/(v_{p1} + v_{p2})$. Each data point represents the average of the expression fraction from 20 to 40 cells. **a** Data from cells expressing RXRLBD-EYFP and EGFP with $P1 = RXRLBD\text{-}EYFP$, $P2 = EGFP$, $p1 = pRXRLBD\text{-}EYFP$, $p2 = pEGFP$. **b** Data from cells expressing RARLBD-EYFP and EGFP with $P1 = RARLBD\text{-}EYFP$, $P2 = EGFP$, $p1 = pRARLBD\text{-}EYFP$, $p2 = pEGFP$. The fit recovered expression efficiency ratios of $y = 0.22$ and $y = 0.19$ for (**a**) and (**b**), respectively. **c** Data from cells expressing RXRLBD-EGFP and EYFP with $P1 = RXRLBD\text{-}EGFP$, $P2 = EYFP$, $p1 = pRXRLBD\text{-}EGFP$, $p2 = pEYFP$ with an expression efficiency ratio of $y = 0.19$. The *insets* show the average protein ratio versus the plasmid mixing ratio for each protein pair with the solid line representing Eq. 7 using the fitted efficiency ratio
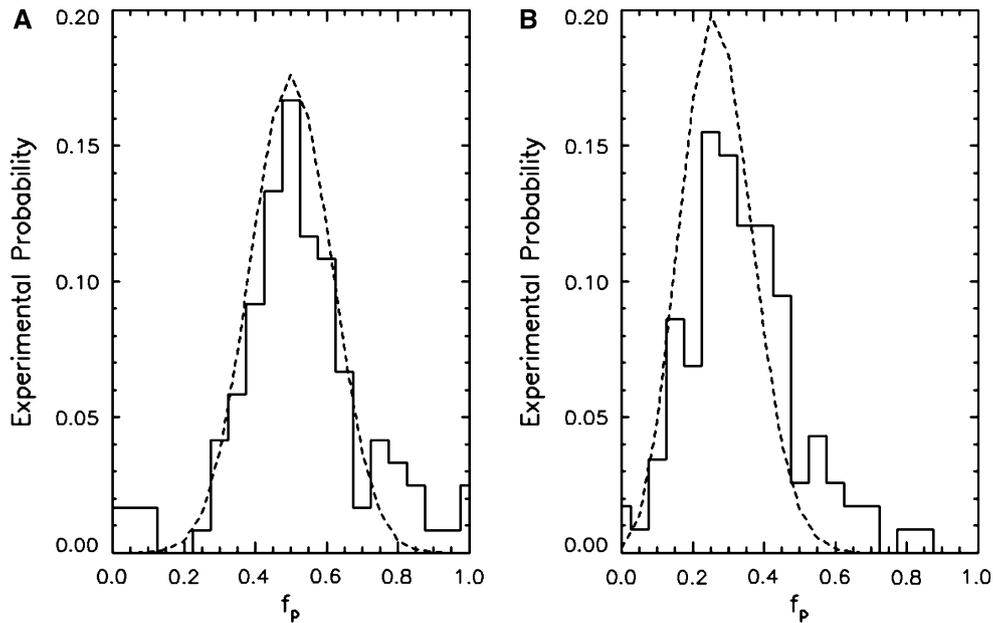
protein expression fraction, $f_p = N_{RXRLBD\text{-}EGFP}/(N_{EYFP} + N_{RXRLBD\text{-}EGFP})$ was determined individually for each measured cell. Figure 2c shows the average protein expression fraction versus plasmid mixing fraction for EYFP and RXRLBD-EGFP. A fit of the data by Eq. 9 determines an efficiency fraction, $y = 0.19$. Note, the data are plotted with the x-axis corresponding to a protein fraction $f_p = N_{RXRLBD\text{-}EGFP}/(N_{EYFP} + N_{RXRLBD\text{-}EGFP})$; plots previously shown all have the EYFP labeled protein in the numerator of the expression fraction. The fit of the data determined $y = 0.19$. The uncertainty in fitting the efficiency ratio y is estimated to be 0.04 for the data of Figs. 1 and 2. Since the efficiency ratios of both data sets (Figs. 2a, c) differ by less than 0.04, we determine that swapping the protein labels has no significant effect on the experiment.

These results demonstrate that the simple linear model (Eq. 7) relating the average protein expression ratio to the pDNA mixing ratio is very successful in describing actual experiments. Again it is important to note that the linear relationship between protein and plasmid ratios leads to a nonlinear relationship for the fractions (Eq. 9), if the efficiency ratio $y \neq 1$.

Variation in protein expression fractions

So far we have focused on the average behavior over a whole cell population. We now take a closer look at the variability from cell to cell. The distribution of the protein expression fractions from a population of cells is best represented by a histogram. Figure 3a shows the histogram of protein expression fractions f_P for cells transfected by a 1:1 pEYFP to pEGFP plasmid ratio. The histogram contains protein expression fraction from 120 cells with a bin resolution of 0.05. The y-axis represents the experimental probability $p(f_P)$ of observing the expression fraction f_P and is normalized by dividing the number of cells $N(f_P)$ with protein expression fraction f_p by the total number of cells sampled. Thus, the histogram directly reflects the experimentally determined probability distribution function (pdf) of the protein expression fraction $\text{pdf}_P(f_p)$. The data are

fraction f_S and the average protein expression fraction $\langle f_P \rangle$ by repeating the measurement of the EGFP/RXRLBD-EYFP pair with switched labels (EYFP/RXRLBD-EGFP). The

Fig. 3 Histogram of protein fractions (*solid line*) for EGFP and EYFP transfected by a pEGFP/pEYFP mixture. The *y-axis* represents the number of cells with a protein expression fraction f_P normalized by the number of cells sampled. A protein fraction of 0 indicates a cell expressing only EGFP, and a value of 1 represents a cell expressing only EYFP. **a** The pEGFP/pEYFP mixture is 1:1. The distribution is

peaked at a protein expression fraction of 0.5 as expected for a 1:1 plasmid ratio. Binomial probability distribution (*dashed curve*) is for $M = 20$ active plasmids and $p_y = 0.5$. **b** The pEGFP/pEYFP mixture is 3:1. The binomial probability distribution (*dashed curve*) is for $M = 20$ active plasmids with $p_y = 0.27$

plotted such that a protein fraction of 0 represents a cell expressing only EGFP, and a value of 1 represents a cell expressing only EYFP. The experimentally measured distribution (solid line) of Fig. 3a is approximately symmetric with respect to its peak at $f_P = 0.5$. These properties of the distribution agree with the expectations for a 50/50% plasmid ratio with an efficiency ratio of $y = 1$.

As pointed out in Eq. 10, the probability distribution of the protein fraction $\mathrm{pdf}_P(f_p)$ is related to the probability distribution function of the active plasmid fraction $\mathrm{pdf}_A(f_A)$ inside the cell. For the special case of $y = 1$, the relationship simplifies to $\mathrm{pdf}_A(f_A) = \mathrm{pdf}_P(f_P)$. In other words, the histogram of Fig. 3a also represents the probability distribution of the active plasmid fraction. Because uptake and activation of a plasmid do not depend on whether it is an EGFP or EYFP plasmid, the probability p_y that a given active plasmid encodes EYFP is expected to be the same as the fraction f_S of EYFP-encoding plasmid in solution, $p_y \approx f_S$.

We now introduce a basic statistical model to describe the distribution of active plasmids. Consider a cell with a total of M active plasmids. The probability that M_y of the M active plasmids are EYFP plasmids is described by a binomial distribution,

$$P_B(M_y, M, p_y) = \binom{M}{M_y} p_y^{M_y} (1 - p_y)^{M - M_y} \tag{12}$$

Let us assume that all cells of the population have M active plasmids. The active plasmid fraction f_A of a given cell is given by $f_A = M_y/M$. Rewriting of Eq. 12 results in a distribution describing the probability distribution of the active plasmid fraction,

$$\mathrm{pdf}_A(f_A, M, p_y) = P_B(M \cdot f_A, M, p_y) \tag{13}$$

Our model predicts that the probability distribution of the active plasmid fraction follows a binomial distribution. The dashed curve in Fig. 3a represents a binomial distribution with $M = 20$ and $p_y = 0.5$. The distribution accurately represents the central peak of the data assuming 20 active plasmids per cell with a 50% probability p_y of a given plasmid being pEYFP. This probability matches well with the EYFP plasmid fraction in solution $f_S = 0.5$. Although the binomial distribution fails to reproduce the tails of the experimental distribution, which are more strongly populated than expected by our model, the binomial distribution provides a reasonable model to estimate the variability of the protein expression ratios, since the population of the tails is sparse. For example, the standard deviation of the experimental protein expression fraction of Fig. 3 is 0.12, while the binomial distribution leads to a standard deviation of 0.11.

We now investigate the distribution function for a plasmid mixing ratio that differs from the symmetric 1:1 ratio used above. Cells were transfected with a 1:3 pEYFP
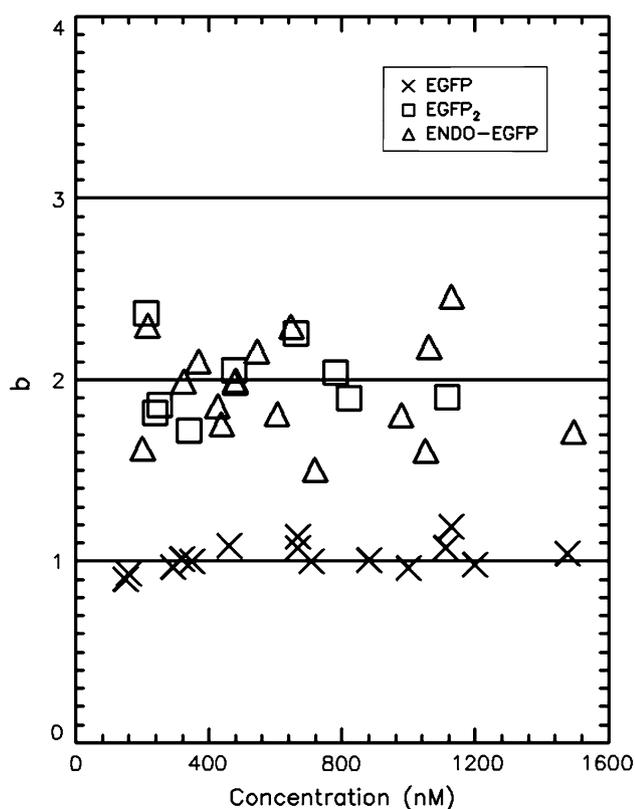
Fig. 4 Normalized brightness b of Endo-EGFP, EGFP, and EGFP$_2$. The brightness of Endo-EGFP is nearly identical to EGFP$_2$, which suggests that Endo-EGFP forms a tight dimer over the concentration range studied. The average normalized brightness for EGFP$_2$ is 1.99 ± 0.03, and the average normalized brightness for Endo-EGFP is 1.95 ± 0.04

to pEGFP plasmid ratio, and the protein expression fraction was determined from 115 cells. The pdf of the protein expression fractions f_p is shown in Fig. 3b. The distribution is asymmetric with a peak shifted to $f_p = 0.25$. The dashed curve in Fig. 3b represents a binomial distribution $P_B(f_p, M, p_y)$ with $M = 20$ and $p_y = 0.27$. Again, the binomial distribution approximately describes the central peak of the distribution, but fails to reproduce the tail of the distribution. The probability for an EYFP-plasmid $p_y = 0.27$ is in good agreement with the 1:3 pEYFP to pEGFP plasmid ratio ($p_S = 0.25$). Our modeling assumed that the number of active plasmids per cell is a constant ($M = 20$), which is unrealistic. The number of active plasmids is expected to be distributed, which needs to be accounted for in a refined model. We will return to this point later in the discussion section.

Application of protein coexpression model to FFS bright and dark experiment

The above experiments relied on two distinctly colored fluorescent proteins to develop a model describing the

statistics of the protein expression fraction from a population of transfected cells. We now apply this model to an FFS study where two proteins are expressed, but only one of the two protein species is labeled. This bright and dark experiment provides a promising approach for identifying the presence of adverse label effects. To illustrate, let us consider the ideal case where the fluorescent label has no influence on protein assembly. In this case, we expect a specific reduction in the measured brightness, because the dark protein is not visible, but competes with the labeled protein in the formation of complexes. For example, an n-meric protein complex with a fraction f_P of labeled proteins and a corresponding fraction $f_{PD} = (1 - f_P)$ of unlabeled proteins leads to a normalized brightness of (Hillesheim et al. 2006)

$$b = n - f_{PD}(n - 1). \tag{14}$$

The normalized brightness $b = \lambda/\lambda_1$, is defined as the ratio between the measured brightness λ of the complex and the brightness λ_1 of the label EGFP, and reflects the oligomerization state of a homo-protein complex (Chen et al. 2005). This property yields a normalized brightness of n for an n-meric complex if all proteins are labeled. Introducing dark proteins reduces the brightness to a value that depends only on the fraction of unlabeled protein. Although the exact value of the unlabeled protein fraction f_{PD} is unknown to us, the model of protein coexpression fractions introduced earlier allows us to predict the distribution and average normalized brightness from a bright and dark experiment. For example, we expect according to Eq. 14 an average normalized brightness of

$$\langle b \rangle = n - \langle f_{PD} \rangle (n - 1) \tag{15}$$

The coexpression model (Eq. 9) relates the average protein fraction $\langle f_{PD} \rangle$ to the fraction of plasmids encoding the dark protein f_S. We test this relationship by conducting an FFS bright and dark experiment on the homo-dimeric protein Endophilin A2.

First, we conduct fluorescence fluctuation spectroscopy (FFS) measurements in the cytoplasm of cells expressing EGFP and the tandem-dimer EGFP$_2$ as a control experiment to confirm brightness doubling (Chen et al. 2002, 2003). The normalized brightness as a function of protein concentration shown in Fig. 4 yields an average normalized brightness of 1.99 ± 0.03 for EGFP$_2$ (Chen et al. 2003). In addition, we measured the brightness of EGFP-labeled Endophilin A2 (Endo-EGFP) as a function of protein concentration (Fig. 5). Our data with an average normalized brightness of 1.95 ± 0.04 confirm the dimeric state of cellular Endophilin A2 as recently shown by Ross et al. (2011). After establishing the dimeric nature of Endophilin A2, a bright and dark experiment with a 50/50% mole ratio of unlabeled (pEndo) and labeled
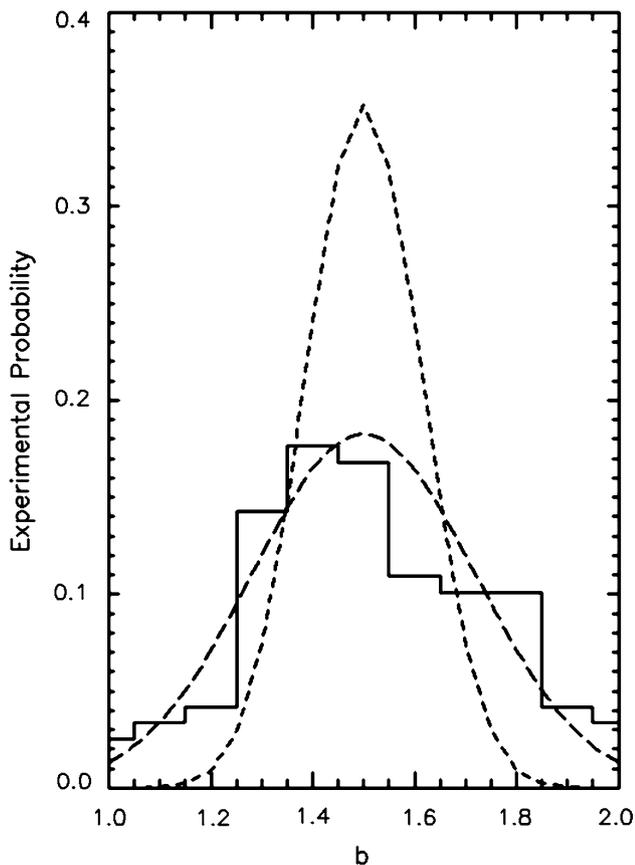
Fig. 5 Histogram of the normalized brightness from cells transfected with a 1:1 pEndo-EGFP/pEndo mixture. The histogram (*solid line*) contains the normalized brightness from 120 cells with a bin resolution of 0.1. The normalized brightness $b = 1$ corresponds to the brightness of monomeric EGFP. The *short dashed curve* represents a binomial distribution for $M = 20$ active plasmids. The convolution of a binomial distribution for $M = 20$ active plasmids with a Gaussian distribution with a standard deviation of 0.19 is shown as a *long dashed line*

(pEndo-EGFP) protein is performed. The normalized brightness values from 120 cells were measured and are plotted in Fig. 5 as a histogram with a bin resolution of 0.1. The average normalized brightness from the population of cell measurements is $\langle b \rangle = 1.51$. This drop in brightness indicates according to Eq. 15 an average fraction for the expressed dark protein of $\langle f_{PD} \rangle = 0.5$, which is identical to the fraction of plasmids encoding the dark protein $f_S = \langle f_{PD} \rangle = 0.5$. This result further implies that the expression efficiency ratio y is one according to Eq. 9. Thus, our data demonstrate that quantitative interpretation of the average brightness from bright and dark experiments provides a measure of the efficiency ratio y. Extending this study to other values of the dark protein-encoding plasmid fraction would provide independent measurements of the efficiency ratio y. A systematic inconsistency in the recovered y values would be indicative of interference by the label.

We now turn to the distribution of the measured brightness values. According to Eq. 14 the brightness is determined by $b = n - f_{AD}(n - 1)$, where f_{AD} is the active plasmid fraction encoding dark protein. For a dimer this reduces to $b = 1 + f_A$, where f_A, the labeled active plasmid fraction, is related to f_{AD} by $f_A = (1 - f_{AD})$. This relationship relates the brightness to the active plasmid distributions, $\mathrm{pdf}(b) = \mathrm{pdf}_A(b - 1)$. Straightforward application of our binomial model (Eq. 13) with $M = 20$ active plasmids and $p_{PD} = f_S = 0.5$ fails to reproduce the measured brightness distribution (short dashed line in Fig. 5). We need to account for the inherent uncertainty of measuring a brightness value inside cells. The standard deviation in brightness calculated from the data in Fig. 5 is 10% for EGFP and EGFP$_2$ and 13% for Endo-EGFP. Thus, we convolute the binomial distribution with a Gaussian distribution representing a relative uncertainty of 13% with respect to the mean brightness of 1.5. The convoluted distribution (long dashed line in Fig. 5) provides a good approximation of the experimental brightness distribution. Thus, dark experiments contain two significant sources of noise, the inherent uncertainty of the brightness measurement and the variations in the dark protein fraction. Both noise sources are independent, and their standard deviations add up as $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$. A 13% uncertainty in the brightness measurement ($\sigma_1 = 0.19$) and the uncertainty due to the active plasmid fraction ($\sigma_2 = 0.11$) lead to a combined uncertainty of $\sigma = 0.22$, which is in good agreement with the standard deviation of 0.25 of the experimental data.

Discussion of general results

It is well known that the protein expression level from transient transfection varies widely from cell to cell. Despite this seemingly random behavior of the absolute expression level, the protein expression ratio of two simultaneously transfected proteins is surprisingly well behaved. We developed a simple model for the average protein expression ratio, its variance, and distribution function. The core of the model is based on the assumption that the vast majority of cellular factors influencing gene expression affect both plasmids in the same way. Thus, these common factors cancel in the ratio of the protein expression, and the only two factors remaining are the expression efficiency ratio y and the active plasmid ratio r_A. Neither of these factors is known a priori, but over a whole cell population, the mean of r_A equals the ratio r_S of plasmids in the transfection solution $\langle r_A \rangle = r_S$. This statement provides a testable hypothesis in the form of Eqs. 7 and 9, because it relates the average protein expression ratio $\langle r_P \rangle$ to the plasmid ratio r_S with y as the only

unknown. We tested this relation by experiment and found that the data agree with the hypothesis. Fitting of the data provides the efficiency ratio y. Further, we describe the cell-to-cell variation in the expression fraction f_P through the introduction of a distribution model of the active plasmid fraction. This was based on a binomial distribution with a total of M active plasmids from both sources, $pDNA_1$ and $pDNA_2$. Clearly, assuming that each cell has a fixed number M of active plasmids is unrealistic. This number is expected to be distributed. In this sense, the value of $M = 20$ used in this study represents an approximation for the mean number of active plasmids.

Despite the simple nature of the model, it very closely approximates the experimental data. This descriptive power of the model should be useful for predicting the feasibility and outcome of many coexpression fluorescence experiments. Our data demonstrate that the mean protein expression ratio can be adjusted in a controlled and predictable fashion by changing the plasmid ratio. This degree of control is useful for selecting conditions that optimize the detection of protein-protein interactions. For example, if one protein expresses much less than the other, this imbalance can be corrected for by adjusting the molar plasmid ratio accordingly. This control over protein coexpression is limited by the finite number of active plasmids present in each cell. The binomial distribution model with $M = 20$ plasmids describes the spread in the experimental data that corresponds to ~ 10 active copies for each type of plasmid in a 1:1 transfection mixture. In trials where pEGFP and pEYFP are mixed at ratios of 1:13 and 1:27, the expressed protein distribution is skewed to be almost entirely EYFP (data not shown). Thus, once protein expression ratios need to be modified by a factor of more than 10, adjusting the pDNA ratio will not produce the desired outcome. This effect was encountered in a study where the large difference in expression efficiency prevented the use of the full-length coactivator SRC-1 (Chen and Müller 2007). This distribution model should be tested for each specific cell line and transfection method used in a study to establish the number of active plasmids, because its value determines the accessible range of expression ratios.

The binomial distribution model is further useful for fluorescence experiments, because it imposes limits on the cell-to-cell variations in brightness values and FRET efficiencies due to different expression ratios. This model predicts the spread in the experimental data and provides a quantitative estimate of the standard deviation of the measured signal. Our experimental data show that a few cells do deviate significantly from the binomial distribution model (Fig. 3). Plotting the coexpression distribution as introduced in this paper is expected to be a useful tool for other studies in order to confirm that the experiment is not selectively sampling cells that deviate from the binomial model.

We applied the cotransfection model to a bright and dark experiment on endophilin A2. The cotransfection model provides a quantitative explanation for the drop in the average brightness. The model further explains the increased scatter in brightness due to variations in the expression ratio. These results demonstrate that quantitative interpretation of FFS bright and dark experiment is feasible once the statistics of coexpression is known. A previous study employed the concept of the bright and dark experiment to check for a labeling effect in the Gag copy number of HIV virus-like particles (Chen et al. 2009). Brightness measurements of virus-like particles containing labeled and unlabeled Gag demonstrated that the fluorescent label did not interfere with the particle assembly process. However, the analysis was based on the assumption that the plasmid ratio is proportional to the expression ratio of labeled and unlabeled Gag. The work presented here justifies this assumption and provides a solid foundation for future bright and dark experiments.

Specific considerations for transient transfection

It is remarkable that the outcome of a complicated process like transient cotransfection is captured by a model that omits the details of the transfection process. In the following we examine the details of cationic lipid-mediated transfection and their influence on the data. Cationic lipids interact with pDNA to form spontaneous complexes called lipoplexes through electrostatic interactions. The overall positive charge of the lipoplex aids in the endocytotic uptake by the cell (El Ouahabi et al. 1997; Friend et al. 1996; Zhou and Huang 1994; Zuhorn et al. 2002). In order for transcription to occur, the lipoplex must escape the endosome, and the pDNA must disassociate from the lipid en route to or inside the nucleus. The number of lipoplexes involved in the dissociation process, the nature of the dissociation process, and the entry into the nucleus are not yet understood (Friend et al. 1996; Kamiya et al. 2002; Brunner et al. 2000). For example, although there is evidence that pDNA fully dissociates from the cationic lipid upon exiting the endosome, there are also conflicting observations (Zhou and Huang 1994; Xu and Szoka Jr 1996; Ouahabi et al. 1999; Merkle et al. 2004; Caracciolo et al. 2009). We are able to use the statistics of expression fractions to contribute to the understanding of some of these details.

Kriess et al. (1999) reported that a single lipoplex contains ~ 37.5 kbase of pDNA. Since the EGFP and EYFP plasmids are 4.5 kbase in length, we estimate that in our study a single lipoplex contains ~ 8 plasmids. Using this as the average number of plasmids per lipoplex, the complete
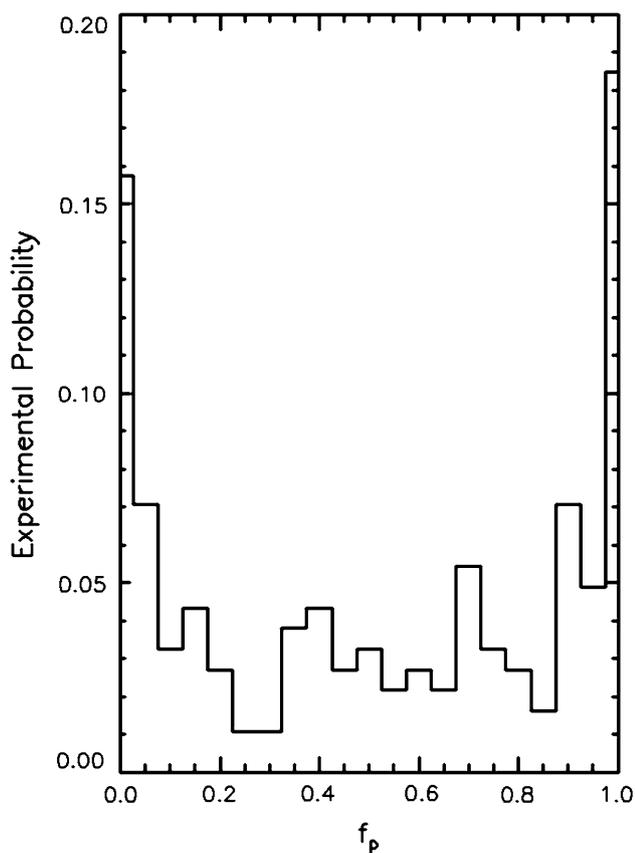
Fig. 6 Histogram of protein fractions for EGFP and EYFP transfected by a 1:1 lipoplex(pEGFP)/lipoplex(pEYFP) mixture. The histogram representing the probability distribution contains protein expression fraction from 140 cells with a bin resolution of 0.05. The *y-axis* represents the number of cells with a protein expression fraction f_P normalized by the number of cells sampled. A protein fraction of 0 represents a cell expressing only EGFP, and a value of 1 represents a cell expressing only EYFP

uncoating of 2–3 lipoplexes is sufficient to provide ~ 20 active plasmids as currently required by our model. Thus, we predict that the number of lipoplexes contributing to protein expression is very small. Since both the number of plasmids per lipoplex and the number of uncoating events are likely distributed, some cells will only experience the uncoating of a single lipoplex. If this lipoplex contains a small number of plasmids, then the probability that they all encode the same protein is much higher than for the bulk of $M = 20$ active plasmids. This scenario provides a plausible explanation for the highly populated tails of the distribution in Fig. 3.

We now test this hypothesis by altering the transfection protocol. So far, our studies involved lipoplexes carrying a statistical mix of two plasmids, but now we switch to a mix of two types of lipoplexes where each type carries the same plasmid. Lipoplexes that carry pEGFP are denoted as pEGFP-lipoplexes, while lipoplexes carrying pEYFP are labeled as pEYFP-lipoplexes. We transfected cells with a

solution containing a 1:1 molar ratio of pEGFP-lipoplexes to pEYFP-lipoplexes. The distribution of protein expression fractions of this experiment (Fig. 6) is distinctly different from Fig. 3. If many lipoplexes are uncoated, the chance that all active plasmids are of the same type is vanishingly small. Such events contribute exclusively to the center of the distribution, which is not representative of the experimental data. On the other hand, the uncoating of a single lipoplex leads to cells expressing only EGFP or EYFP which populate the tails of the distribution ($f_P = 0$ and $f_P = 1$). While this model reproduces the tails of the experimental data, we also need to include the uncoating of two to a few lipoplexes in order to account for the middle of the distribution representing the expression of both proteins. Thus, our experiment convincingly demonstrates that very few lipoplexes contribute to protein expression. This discussion provides some additional insight. First, because although the total amount of pDNA (active and inactive forms) present in a transfected cell is very high (Bengali et al. 2009; Hama et al. 2006, 2007), the dissociation of active plasmid from the complex is a very rare event. Second, our cotransfection model, although it does not account for the existence of lipoplexes, is adequate provided plasmids are mixed together before adding the lipid reagent.

Conclusions

Our model provides a novel quantitative characterization of the average and standard deviation of protein coexpression ratios within living cells. With such a model, it is now possible to predict and control relative protein expression through manipulation of pDNA ratios prior to transfection, which is of interest for fluorescence studies of protein-protein interactions and for FFS bright and dark experiments. The distribution model also provides a prediction for variations in the fluorescence signal due to different expression ratios. Additional studies using different cell lines and other transfection protocols will be needed to test the robustness and generality of our findings. It is worthwhile to again point out that all the plasmids of this study use the same promoter. It would be interesting to repeat the study for plasmids with different promoters and check if the model still applies by using a modified expression efficiency ratio. Nevertheless, the model of expression fractions established in this article provides insight into the cationic lipid-mediated transfection process and should serve as a quantitative platform for future studies of protein coexpression and fluorescence applications in cells.

References

Bengali Z, Rea JC, Gibly RF, Shea LD (2009) Efficacy of immobilized polyplexes and lipoplexes for substrate-mediated gene delivery. Biotechnol Bioeng 102:1679–1691. doi:10.1002/bit.22212

Berland KM (2004) Fluorescence correlation spectroscopy: a new tool for quantification of molecular interactions. Methods Mol Biol 261:383–398. doi:10.1385/1-59259-762-9:383

Brunner S, Sauer T, Carotta S et al (2000) Cell cycle dependence of gene transfer by lipoplex, polyplex and recombinant adenovirus. Gene Ther 7:401–407

Caracciolo G, Caminiti R, Digman MA et al (2009) Efficient escape from endosomes determines the superior efficiency of multi-component lipoplexes. J Phys Chem B 113:4995–4997

Chen Y, Müller JD (2007) Determining the stoichiometry of protein heterocomplexes in living cells with fluorescence fluctuation spectroscopy. Proc Natl Acad Sci USA 104:3147–3152. doi:10.1073/pnas.0606557104

Chen Y, Müller JD, Ruan Q, Gratton E (2002) Molecular brightness characterization of EGFP in vivo by fluorescence fluctuation spectroscopy. Biophys J 82:133–144. doi:16/S0006-3495(02)75380-0

Chen Y, Wei L-N, Müller JD (2003) Probing protein oligomerization in living cells with fluorescence fluctuation spectroscopy. Proc Natl Acad Sci USA 100:15492–15497. doi:10.1073/pnas.2533045100

Chen Y, Wei L-N, Müller JD (2005) Unraveling protein–protein interactions in living cells with fluorescence fluctuation brightness analysis. Biophys J 88:4366–4377. doi:10.1529/biophysj.105.059170

Chen Y, Wu B, Musier-Forsyth K et al (2009) Fluorescence fluctuation spectroscopy on viral-like particles reveals variable gag stoichiometry. Biophys J 96:1961–1969. doi:16/j.bpj.2008.10.067

Ciruela F (2008) Fluorescence-based methods in the study of protein–protein interactions in living cells. Curr Opin Biotechnol 19:338–343

Day RN, Davidson MW (2009) The fluorescent protein palette: tools for cellular imaging. Chem Soc Rev 38:2887–2921. doi:10.1039/b901966a

Day RN, Schaufele F (2008) Fluorescent protein tools for studying protein dynamics in living cells: a review. J Biomed Opt 13:031202

El Ouahabi A, Thiry M, Pector V et al (1997) The role of endosome destabilizing activity in the gene transfer process mediated by cationic lipids. FEBS Lett 414:187–192

Felgner PL, Gadek TR, Holm M et al (1987) Lipofection: a highly efficient, lipid-mediated DNA-transfection procedure. Proc Natl Acad Sci USA 84:7413–7417

Friend DS, Papahadjopoulos D, Debs RJ (1996) Endocytosis and intracellular processing accompanying transfection mediated by cationic liposomes. Biochim Biophys Acta 1278:41–50

Gambin Y, Deniz AA (2010) Multicolor single-molecule FRET to explore protein folding and binding. Mol Biosyst 6:1540–1547. doi:10.1039/c003024d

Hama S, Akita H, Ito R et al (2006) Quantitative comparison of intracellular trafficking and nuclear transcription between adenoviral and lipoplex systems. Mol Ther 13:786–794

Hama S, Akita H, Iida S et al (2007) Quantitative and mechanism-based investigation of post-nuclear delivery events between adenovirus and lipoplex. Nucleic Acids Res 35:1533–1543. doi:10.1093/nar/gkl1165

Haustein E, Schwille P (2007) Fluorescence correlation spectroscopy: novel variations of an established technique. Annu Rev Biophys Biomol Struct 36:151–169. doi:10.1146/annurev.biophys.36.040306.132612

Hillesheim LN, Chen Y, Müller JD (2006) Dual-color photon counting histogram analysis of mRFP1 and EGFP in living cells. Biophys J 91:4273–4284. doi:29/biophysj.106.085845

Kamiya H, Fujimura Y, Matsuoka I, Harashima H (2002) Visualization of intracellular trafficking of exogenous DNA delivered by cationic liposomes. Biochem Biophys Res Commun 298:591–597

Kreiss P, Cameron B, Rangara R et al (1999) Plasmid DNA size does not affect the physicochemical properties of lipoplexes but modulates gene transfer efficiency. Nucleic Acids Res 27:3792–3798

Ma Z-L, Werner M, Körber C et al (2007) Quantitative analysis of cotransfection efficiencies in studies of ionotropic glutamate receptor complexes. J Neurosci Res 85:99–115. doi:10.1002/jnr.21096

Merkle D, Lees-Miller SP, Cramb DT (2004) Structure and dynamics of lipoplex formation examined using two-photon fluorescence cross-correlation spectroscopy. Biochemistry 43:7263–7272

Müller JD (2004) Cumulant analysis in fluorescence fluctuation spectroscopy. Biophys J 86:3981–3992

Ouahabi AE, Thiry M, Schiffmann S et al (1999) Intracellular visualization of BrdU-labeled plasmid DNA/cationic liposome complexes. J Histochem Cytochem 47:1159–1166. doi:10.1177/002215549904700908

Ross JA, Chen Y, Muller J et al (2011) Dimeric endophilin A2 stimulates assembly and GTPase activity of dynamin 2. Biophys J 100:729–737

Snapp EL (2009) Fluorescent proteins: a cell biologist's user guide. Trends Cell Biol 19:649–655. doi:16/j.tcb.2009.08.002

Voges D, Watzele M, Nemetz C et al (2004) Analyzing and enhancing mRNA translational efficiency in an Escherichia coli in vitro expression system. Biochem Biophys Res Commun 318:601–614

Xu Y, Szoka FC Jr (1996) Mechanism of DNA release from cationic liposome/DNA complexes used in cell transfection. Biochemistry 35:5616–5623

Zhou X, Huang L (1994) DNA transfection mediated by cationic liposomes containing lipopolylysine: characterization and mechanism of action. Biochim Biophys Acta 1189:195–203

Zuhorn IS, Kalicharan R, Hoekstra D (2002) Lipoplex-mediated transfection of mammalian cells occurs through the cholesterol-dependent clathrin-mediated pathway of endocytosis. J Biol Chem 277:18021–18028