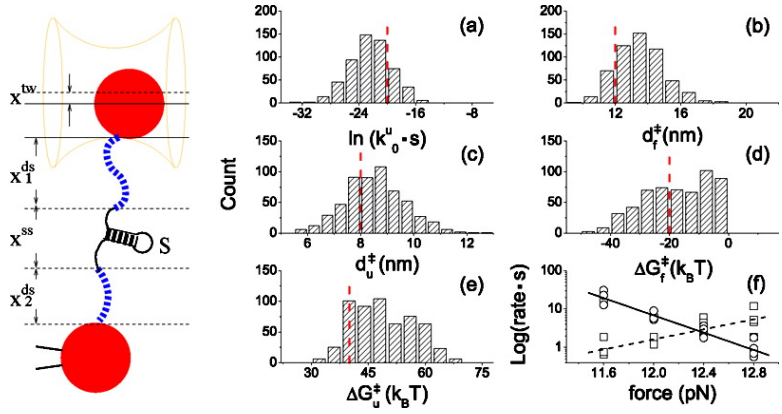
THE JOURNAL OF PHYSICAL CHEMISTRY B

Letter

Bayesian Analysis of Folding and Unfolding Time Series of Single-Forced RNAs

Xiao Chuan Xue, Huan Tong, Fei Liu, and Zhong-can Ou-Yang

J. Phys. Chem. B, **2008**, 112 (44), 13680-13683• DOI: 10.1021/jp8020886 • Publication Date (Web): 09 October 2008



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

The Journal of Physical Chemistry B is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

THE JOURNAL OF PHYSICAL CHEMISTRY B LETTERS

2008, *112,* 13680–13683
Published on Web 10/09/2008

Bayesian Analysis of Folding and Unfolding Time Series of Single-Forced RNAs

Xiao Chuan Xue,[†] Huan Tong,[†] Fei Liu,*[,†] and Zhong-can Ou-Yang[†,‡]

Center for Advanced Study, Tsinghua University, Beijing, 100084, and Institute of Theoretical Physics, The Chinese Academy of Sciences, P.O. Box 2735 Beijing 100080, China

Received: March 10, 2008; Revised Manuscript Received: August 27, 2008

On the basis of a coarse-grain physical model of the folding and unfolding of single-forced RNAs conducted in light tweezer experiments, we theoretically investigate the feasibility of inferring the RNA's intrinsic kinetic parameters from the noisy time series of the molecule's extension. A Bayesian approach using Monte Carlo Markov Chain is proposed. We prove that this statistical approach is efficient and accurate in inferring the molecule's physical parameters, even if the experimental data are yielded under a narrow range of forces.

Introduction

The current single-molecule manipulation provides a novel approach to study the kinetics of single RNAs. Different from many conventional experimental techniques, such as X-ray crystallography, which usually only provide static pictures of the molecules, the current manipulation techniques, mainly including optical tweezer, can trace the full folding/unfolding processes of one RNA by monitoring its extension or force exerted on the molecule in real time.[1−3]

As many nano- or mesoscopic systems, the behavior of single RNAs (∼30 nm) in a light tweezer is highly dynamic and noisy. In practice, the situation becomes more complicated; in order to manipulate one RNA by the optical trapping method, the RNA must first be tethered between two large dielectric beads (∼μm) through two long double-stranded DNA/RNA handles (∼μm); see Figure 1. The folding/unfolding of single-forced RNAs could be conducted in two types of manipulation experiments. One is the constant force mode (CFM), where the experimental control parameter, a constant force F of preset value, is applied on the bead in the light tweezer with or without feedback control.[1−3] The other is the passive mode (PM), where the control parameter, the distance between the centers of the light tweezer and the bead held by the micropipette, $x_T = x^{tw} + x_1^{ds} + x^{ss} + x_2^{ds}$, is left stationary[3] (the sizes of the beads are not included for they do not matter in our discussion). Due to the presence of the beads and handles, it would be expected that the kinetics of the RNA in the light tweezer experiment is distinct from the kinetics of the linker-free RNA. Hence, how to extract the intrinsic kinetic information of single RNAs from the experimental data is an intriguing biophysical issue. One possible strategy is to find optimal experimental conditions through experimental comparison and computational simulation.[3,4] An alternative way is to collect the existing RNA kinetic data and infer the intrinsic parameters by advanced statistic approaches. To the best of our knowledge, the latter was not quantitatively implemented in the literature. In this Letter, we present such an effort.

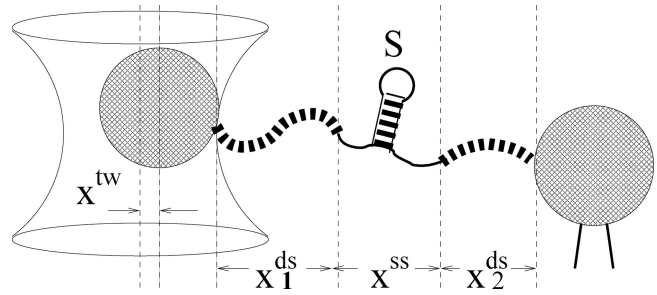


Figure 1. Sketch of the folding/unfolding of a forced RNA in a light tweezer. The RNA molecule is attached between the two beads (the big shaded points) with two long DNA/RNA hybrid handles (the black dashed curves).

Physical Model

The system shown in Figure 1 involves several time scales: the relaxation time of the bead in the tweezer, τ_b, the relaxation time of the handles and single-stranded (ss) RNA, τ_h and τ_{ssRNA}, the characteristic time of the folding/unfolding kinetics of the RNA, τ_{f-u}, and the characteristic time of the opening/closing of single base pairs, τ_{bp}.[4,5] Under the conventional experimental conditions,[1−3] the relaxation times τ_h, τ_{ssRNA}, and τ_{bp} are always far shorter than the relaxation times of the bead and folding/unfolding RNA kinetics.[4,5] Hence, it is plausible to assume that the extension of the handles and ssRNA are in thermal equilibrium instantaneously.

On the other hand, the experiment has observed that some RNAs can hop between folded (f) and unfolded (u) states when the applied external force closes to the force at which the equilibrium constant of the folding/unfolding "reaction" is equal to 1.[1] In a certain range of forces, a two-state model with folding and unfolding rates seems to be natural to describe the kinetics of these RNAs. However, the fluctuating force, which directly exerts on the molecules and is induced by the position fluctuation

* To whom correspondence should be addressed. E-mail address: liufei@tsinghua.edu.cn.
† Tsinghua University.
‡ The Chinese Academy of Sciences.

10.1021/jp8020886 CCC: $40.75

of the bead in the light tweezer, makes the calculation of the two rates complex. A similar issue has been studied in the resonant activation.[6,7] One important conclusion there is that, if the fluctuation is small and fast, rates are only determined by average free-energy surfaces.[7] This situation is satisfied in the folding/unfolding experiment[3] because the relaxation time of the bead τ_b is mostly shorter than the relaxation time τ_{f-u} of the RNA kinetics. Hence, when the RNA stays in the unfolded (folded) state and no "reaction" occurs, the folding (unfolding) rate is a function of the mean position of the bead.

On the basis of the above discussion, we describe the composite system by two coupled diffusion−reaction equations

$$\frac{\partial}{\partial t}P_f(x,t) = [\mathcal{L}_f - k^u(\langle x\rangle_f)]P_f + k^f(\langle x\rangle_u)P_u$$
$$\frac{\partial}{\partial t}P_u(x,t) = [\mathcal{L}_u - k^f(\langle x\rangle_u)]P_u + k^u(\langle x\rangle_f)P_f \quad (1)$$

where $P_i(x, t)$ (i = f or u) is the joint probability density at the distance $x = x_1^{ds} + x^{ss} + x_2^{ds}$ at time t with the RNA at state i and k^i is the folding or unfolding rate, which is the function of the mean distance $\langle x\rangle_i$ when the RNA is correspondingly at state i = u or f. The Fokker−Planck operators $\mathcal{L}_i$ in the above equations are

$$\mathcal{L}_i = D\frac{\partial}{\partial x}e^{-\beta V_i(x)}\frac{\partial}{\partial x}e^{\beta V_i(x)} \quad (2)$$

where D is the diffusion coefficient, $\beta^{-1} = k_BT$ with k_B as Boltzmann's constant and T as the absolute temperature, $V_i(x)$ is the RNA state-dependent potential and defined as

$$V_i(x) = W_{ext}(x) + \int_0^x f_i(x')dx'$$

with $f_i(x) = [0.25(1 - x/l_i)^{-2} + x/l_i - 0.25]/\beta P_{eff}^i$,[8,9] with the persistent length as P_{eff}^i [10] and the contour length as $l_i = 2L_h + L_{ssRNA}^i$; the external work $W_{ext}(x)$ done by the external force is Fx in the CFM and $\varepsilon(x_T - x)^2/2$ with a tweezer stiffness ε in the PM, respectively. Apparently, the mean distances $\langle x\rangle_i = \int x' p_i^{eq}(x')dx'$, where

$$p_i^{eq}(x) = \frac{\exp[-\beta V_i(x)]}{\int \exp[-\beta V_i(x')]dx'} \quad (3)$$

A number of rate models have been proposed to describe biomolecules' unfolding or rupture processes.[11−13] In this work, we use a new one that is exactly solved under an assumption that the RNA's extension is a good reaction coordinate and the folding and unfolding free-energy surfaces are linear with respect to the coordinate

$$k^u(\langle x\rangle_f) = k_0^u\frac{B[\beta\Delta G_u^{\ddagger}]}{B[\beta(\Delta G_u^{\ddagger} - f_f\langle x\rangle_f d_u^{\ddagger})]} \quad (4)$$

and

$$k^f(\langle x\rangle_u) = k_0^u\left(\frac{d_u^{\ddagger}}{d_f^{\ddagger}}\right)^2\frac{B[\beta\Delta G_u^{\ddagger}]}{B[\beta(\Delta G_f^{\ddagger} + f_u\langle x\rangle_u d_f^{\ddagger})]} \quad (5)$$

where $B(x) = (e^x - x - 1)/x^2$, $\Delta G_i^{\ddagger}$ and $d_i^{\ddagger}$ are, respectively, the barrier heights and the transition distances between the free-energy extremum and the transition state, and k_0^u is the intrinsic unfolding rate of the RNA in the absence of force. Their derivations and the reasons using the linear approximation of the free-energy surface can be found in the Supporting Information.

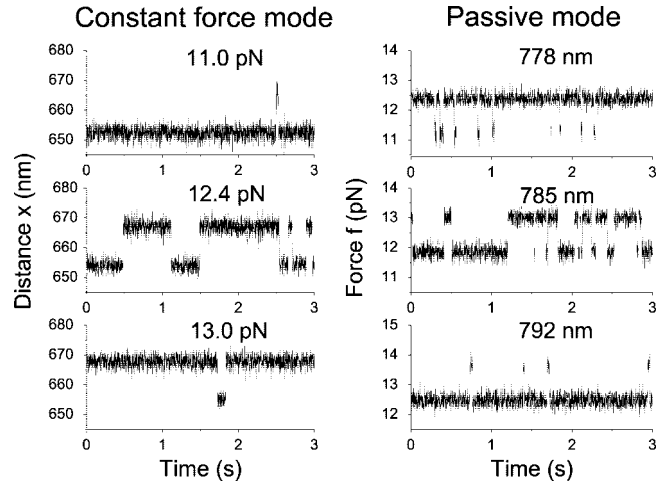


Figure 2. Time series of the distance x at three different constant forces in the CFM (left column) and at three different x_T in the PM (right column). The duration of them is 3 s, and the time interval is 1 ms.

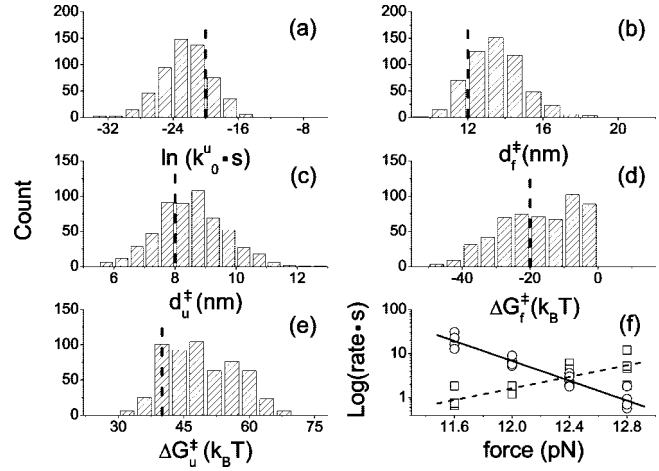


Figure 3. (a−e) Posterior sampling histograms of the five parameters for a data set generated by simulating eq 1 in the CFM. The vertical dashed lines represent the actual values of the parameters. (f) The folding (the circles) and unfolding (the squares) rates evaluated by the HF method from the five data sets. The lines are the fits of the Bell rate model (see the Supporting Information).

Equation 1 has an exact solution under steady state

$$P_i^{ss}(x) = \pi_i p_i^{eq}(x) \quad (6)$$

where $\pi_i = k^i/k$ and $k = k^u + k^f$. This equation is also useful in studying time-dependent cases. As an illustration, we simulate several time series of the distance x at three different values of the control parameters in the CFM and PM; see Figure 2. The simulation parameters are ε = 0.1 pN/nm for the tweezer stiffness, $R_b = 1.0\ \mu$ m for the bead radius, $\eta = 10^{-3}$ kg/ms for the viscosity of water, L_h = 340.0 nm (1000 base pairs) and P_h = 53.0 nm for the contour and persistence lengths of the handle, L_{ssRNA}^u = 20.1 nm (34 bases) and P_{ssRNA} = 1.0 nm for the complete unfolded RNA, L_{ssRNA}^f = 2.0 nm for the folded RNA, $\ln(k_0^u \cdot s) = -20.7$ for the logarithms of the intrinsic unfolding rate, $\Delta G_u^{\ddagger} = 40\ k_BT$ and $\Delta G_f^{\ddagger} = -20\ k_BT$ for the barrier heights, and $d_u^{\ddagger} = 8$ nm and $d_f^{\ddagger} = 12$ nm for the locations of the transition state; all values are in the experimental ranges.[2,3] We see that the simulations are qualitatively consistent with the experimental observation.[3] In the following, we focus our attention on the inference of the intrinsic kinetic parameters from the time series obtained by the simulation.

TABLE 1: A Comparison between the Actual Values and the Mean and MAP Values of the Five Kinetic Parameters Inferred by Our Bayesian Approach in the CFM and PM

		$\ln(k_0^{\mathrm{u}}\cdot\mathrm{s})$	$d_{\mathrm{u}}^{\ddagger}$ (nm)	$\Delta G_{\mathrm{u}}^{\ddagger}$ ($k_{\mathrm{B}}T$)	$d_{\mathrm{f}}^{\ddagger}$ (nm)	$\Delta G_{\mathrm{f}}^{\ddagger}$ ($k_{\mathrm{B}}T$)
	actual value	−20.7	8.0	40.0	12.	−20.0
CFM	mean	-21.3 ± 4.1	8.2 ± 1.5	43.0 ± 8.6	13.4 ± 2.2	-22.1 ± 13.5
	MAP	−22.0	8.5	39.0	12.5	−25.0
PM	mean	-21.4 ± 3.8	8.3 ± 1.4	41.7 ± 6.3	13.3 ± 2.5	-23.2 ± 14.3
	MAP	−21.0	7.5	37.0	11.5	−20.0

Bayesian Parameter Estimates

Let $\mathbf{x} = (x_0, ..., x_n)$ be a sequence of the distances x_l observed at equal separated time intervals t_l at a given constant force F or x_{T}. According to Bayes' theorem, the posterior distribution on the parameters $\theta = [\ln(k_0^{\mathrm{u}} \cdot \mathrm{s}), \Delta G_{\mathrm{f}}^{\ddagger}, \Delta G_{\mathrm{u}}^{\ddagger}, d_{\mathrm{f}}^{\ddagger}, d_{\mathrm{u}}^{\ddagger}]$ given observation $\mathbf{x}$ is

$$P(\theta|\mathbf{x}) \propto \eta(\theta)L(\mathbf{x}|\theta) \tag{7}$$

where $\eta(\theta)$ and $L(\mathbf{x}|\theta)$ are the prior distribution on the parameters and the likelihood function of observing $\mathbf{x}$ given the parameters, respectively. The RNA molecule is either folded or unfolded at any time. Because the light tweezer experiment only records the distance between the centers of the two beads, the folding/unfolding of RNA is virtually a hidden Markov process.[14] Then, the likelihood is

$$L(\mathbf{x}|\theta) = \mathbf{1}^T \times \prod_{l=n}^{1} \mathbf{P}(x_l, t_l|x_{l-1}, t_{l-1}) \times \mathbf{P}_0(x_0) \tag{8}$$

The matrix element $[\mathbf{P}(x, \Delta t|y, 0)]_{\mathrm{ij}}$ (i,j = u,f) in the above equation represents the transition probabilities of eq 1 that observe the RNA state i and the bead position x after a time $\Delta t = t_{l+1} - t_l$ when the current RNA state is j and the position is y. We have assumed the observation starting from the steady-state $\mathbf{P}_0(x_0) = [P_{\mathrm{f}}^{\mathrm{ss}}(x_0), P_{\mathrm{u}}^{\mathrm{ss}}(x_0)]^T$. Equation 1 does not have exact time-dependent solutions. However, considering that in the real experiments the relaxation time of the bead in the light tweezer τ_{b} is mostly shorter than the measurement time Δt, we can safely approximate

$$\mathbf{P}(x, \Delta t|y, 0) \simeq \mathbf{\Lambda}(x)\mathbf{Q}(\Delta t) \tag{9}$$

where

$$\mathbf{\Lambda}(x) = \mathrm{diag}[p_{\mathrm{f}}^{\mathrm{eq}}(x), p_{\mathrm{u}}^{\mathrm{eq}}(x)] \tag{10}$$

and

$$\mathbf{Q}(\Delta t) = \begin{bmatrix} \pi_{\mathrm{f}} + \pi_{\mathrm{u}}\mathrm{e}^{-\Delta t k} & \pi_{\mathrm{f}}(1 - \mathrm{e}^{-\Delta t k}) \\ \pi_{\mathrm{u}}(1 - \mathrm{e}^{-\Delta t k}) & \pi_{\mathrm{u}} + \pi_{\mathrm{f}}\mathrm{e}^{-\Delta t k} \end{bmatrix} \tag{11}$$

It is independent of the initial position y of the bead. With eqs 10 and 11, the likelihood function can be calculated by the forward recursion and ongoing scaling techniques.[14] In order to have sufficient data to make reliable estimates of the parameters, we use multiple sequences obtained at different values of the control parameters, that is, different constant forces F in the CFM or distances x_{T} in the PM. The joint likelihood is simply a multiplication of eq 7 at a certain force or distance. Finally, we choose independent flat priors for the parameters in θ. Because we are treating the logarithm of the rate, its flat prior is equivalent to the Jeffreys' prior[15] of the rate itself.

Direct computation from $P(\theta|\mathbf{x})$ is infeasible. We use the standard Metropolis Monte Carlo algorithm[15] to sample from it. Figure 3a−e illustrates the posterior sampling distributions on the five parameters from a data set in the CFM. We see that all of the distributions are a highly asymmetric, and some of them have long tails, particularly for $\Delta G_{\mathrm{f}}^{\ddagger}$. To exclude the randomness of the choice of data set, we generate another four data sets for the each manipulation mode; each data set is composed of seven time series, which are simulated at seven constant forces (10.4−12.8 pN) or at seven x_{T}'s (776−794 nm); the during time is 12 s and $\Delta t = 1$ ms. We list both the mean and maximum a posteriori (MAP) values of the five parameters inferred from the five data sets in Table 1. We see that in both of the modes, the parameter estimates given by the two statistics agree with the actual values satisfactorily.

Discussion and Conclusion

Compared with the widly used histogram fitting (HF) method (see the Supporting Information), the main advantages of the Bayesian approach developed in this work are as follows. First, the Bayesian approach based on the time intervals can exploit the time series that are yielded under a wilder range of forces. In order to reliably measure the RNA folding/unfolding kinetics, the HF method needs enough hopping events to construct rational dwell time histograms. Due to the stability restriction of the current instrument, this requirement is only satisfied when the experiments are conducted under a very narrow range of forces.[1,3,16] Hence, the kinetic information underlying the time series under the forces out of the narrow range is lost. Figure 3f shows such a case, where the HF method fails to estimate the rates at the forces smaller than 11.6 pN. Second, the Bayesian approach using Monte Carlo Markov Chain is able to accurately and robustly estimate all of the physical parameters, including the two energy barriers, and even the time series are yielded under a small range of forces, for example, ∼2 pN for the five data sets in the CFM. This point is unique. As a demonstration, we attempt to fit eq 4 to the unfolding rates shown in Figure 3f directly and find that the fit always fails to predict $\Delta G_{\mathrm{u}}^{\ddagger}$. This would be expected in that, for the traditional fitting methods such as the least-squares fitting, the force range in Figure 3f is too narrow. The satisfied inference about the intrinsic barriers by our approach is important; in addition to the unfolding barrier being indispensable to yield other reliable unfolding parameters,[12] it may imply a new way to reconstruct the intrinsic free-energy surface by the hopping experiment. Finally, a small number of data are needed for the Bayesian approach; the existing hopping experiments can readily provide this. Our next step is to apply the Bayesian approach to the real experimental data.

Acknowledgment. F.L. would like to thank Drs. Hu Chen and Jie Yan for generously showing us their unpublished calculation about the effective persistence of a sequence of heterogeneous WLCs. This work was supported, in part, by the Tsinghua Basic Research Foundation and by the National Science Foundation of China under Grant No. 10704045.

Supporting Information Available: The derivations of eqs 4 and 5, a discussion about the linear approximation of the free-

energy surface, and a comparison between kinetic parameters inferred by the HF and the Bayesian approaches. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Liphardt, J. B.; Onoa, B.; Smith, S. B.; Tinoco, I., Jr.; Bustamante, C. *Science* **2001**, *292*, 733.

(2) Woodside, M. T.; Anthony, P. C.; Behnke-Parks, W. M.; Larizadeh, K.; Travers, K.; Herschlag, D.; Block, S. M. *Science* **2006**, *314*, 1001.

(3) Wen, J. D.; Manosas, M.; Li, P. T. X.; Smith, S. B.; Bustamante, C.; Ritort, F.; Tinoco, I., Jr. *Biophys. J.* **2007**, *92*, 2996.

(4) Manosas, M.; Wen, J. D.; Li, P. T. X.; Smith, S. B.; Bustamante, C.; Tinoco, I., Jr.; Ritort, F. *Biophys. J.* **2007**, *92*, 3010.

(5) Manosas, M.; Ritort, F. *Biophys. J.* **2005**, *88*, 3224.

(6) Doering, C. R.; Gadoua, J. C. *Phys. Rev. Lett.* **1992**, *69*, 2318.

(7) Pechukas, P.; Hanggi, P. *Phys. Rev. Lett.* **1994**, *73*, 2772.

(8) Marko, J. F.; Siggia, E. D. *Macromolecules* **1995**, *28*, 8759.

(9) Bustamante, C.; Marko, J. F.; Siggia, E. D.; Smith, S. *Science* **1994**, *264*, 1599.

(10) We do not need to model the handles and ssRNA chain independently because the effective persistent length of a sequence of connected worm-like chains (WLCs) can be calculated by the following formula: $P_{\mathrm{eff}} = [(L_1 + . . + L_n)/(L_1/\sqrt{P_1} + . . + L_n/\sqrt{P_n})]^2$, where L_i and P_i ($i = 1,. ., n$) are the contour lengths and persistent lengths of the WLCs, respectively. (Chen and Yan, personal communications).

(11) Bell, G. I. *Science* **1978**, *200*, 618.

(12) Dudko, O. K.; Hummer, G.; Szabo, A. *Phys. Rev. Lett.* **2006**, *96*, 108101.

(13) Lin, H. J.; Chen, H. Y.; Sheng, Y. J.; Tsao, H. K. *Phys. Rev. Lett.* **2007**, *98*, 088304.

(14) Rabiner, L. R. *Proc. IEEE* **1989**, *77*, 257.

(15) Gelman, A.; Carlin, J. B.; Stern, H. S.; Rubin, D. B. *Bayesian Data Analysis*; Chapman and Hall: New York, 1995.

(16) Li, P. T.; Collin, D.; Smith, S. B.; Bustamante, C.; Tinoco, I., Jr. *Biophys. J.* **2006**, *90*, 250.

JP8020886