Detection of Transient Events in the Presence of Background Noise

Wilfried Grange,*,†,‡ Philippe Haas,‡,§ Andreas Wild,‡,§ Michael Andreas Lieb,§ Michel Calame,§ Martin Hegner,† and Bert Hecht*,∥

Centre for Research on Adaptive Nanostructures and Nanodevices, Trinity College Dublin, College Green, Dublin 2, Ireland, National Center of Competence for Research in Nanoscale Science Institute of Physics, University of Basel, Klingelbergstrasse 82, CH-4056 Basel Switzerland, and Nano-Optics and Biophotonics Group, Wilhem Conrad Röntgen Research Center for Complex Materials Systems (RCCM), Department of Experimental Physics 5, University of Würzburg, Germany

Received: March 5, 2008; Revised Manuscript Received: March 5, 2008

We describe a method to detect and count transient burstlike signals in the presence of a significant stationary noise. To discriminate a transient signal from the background noise, an optimum threshold is determined using an iterative algorithm that yields the probability distribution of the background noise. Knowledge of the probability distribution of the noise then allows the determination of the number of transient events with a quantifiable error (wrong-positives). We apply the method, which does not rely on the choice of free parameters, to the detection and counting of transient single-molecule fluorescence events in the presence of a strong background noise. The method will be of importance in various ultra sensing applications.

Introduction

The discrimination of rare transient events (bursts) above a strong stationary background noise with a high level of confidence is a problem of broad importance in various sensing applications ranging from ultrasensitive optical detection[1–3] in biological assays[4,5] or medical diagnostics[6,7] to electromagnetic sensors[8,9] or defense applications.[7] In general, transient signals are considered detectable either if (i) their amplitude is many standard deviations above the mean value of the noise's probability distribution and has a narrow distribution or if (ii) the waveform (the duration of the transient event) is clearly distinct from the noise's characteristic fluctuations in time. Here, we propose a method, which is applicable if the signal bursts are neither large in amplitude nor easily distinguishable from the characteristic fluctuations of the noise. The method is based on a fast converging iterative algorithm, which determines an optimum threshold for the detection and counting of bursts. It provides a user-definable quantitative measure for the probability of false positive events due to the background noise peaks. The reliability of the method is demonstrated by Monte Carlo simulations of the burst detection process. To highlight the method's potential, we detect and count single-molecule fluorescence bursts recorded in the presence of a significant stationary background noise.

Results

We consider a data set describing a time series of counts per time interval containing rare transient events (bursts) above a significant background noise with a Poissonian distribution. Apart from being sufficiently rare, no further assumptions are made with respect to the amplitude and shape distribution of the transient events superimposed to the background noise. Note that the results presented here are applicable for any type of background distribution as long as its shape is known. Figures 1a and b show as an example a data set representing a time trace of single-molecule fluorescence bursts as well as the respective histogram $H(n)$. Here, n is the number of counts per 100 μs time interval (bin). Fluorescence bursts of various amplitudes are observed above the background noise. Consequently $H(n)$ shows a clearly distinguishable main Poissonian noise peak and a tail that accounts for the fluorescence bursts.[3] Signal bursts cannot be fully separated from the background noise since both distributions apparently overlap. To optimally discriminate signal bursts from similar events due to background noise, a threshold must be determined above which a fluctuation is counted as a signal burst. The threshold must on one hand be low enough to miss as few as possible true signal bursts, and on the other hand, it must be high enough to minimize the probability of counting a strong fluctuation of the noise as a signal burst. Wrongly assigned bursts contribute to false positive events, which—in view of applications e.g. in medical diagnostics—must be avoided or at least kept to a quantifiable error.

To determine such an optimum threshold, the probability distribution of the background has to be recovered. Considering the normalized Poissonian distribution $P(n)$, we have

$$P(n) = \frac{e^{-\mu}\mu^n}{n!} \tag{1}$$

where μ is the mean, and $\sigma = \sqrt{\mu}$ the standard deviation.

Assuming that $P(n)$ can be recovered with some degree of accuracy, we may consider the probability distribution of the background alone. This then enables us to determine a threshold for the burst amplitude, χ, by demanding that the absolute number of time intervals K for which the number of counts n exceeds the threshold χ is smaller than a tolerable small number, say α. $K(\chi)$ is determined as

* Corresponding authors. e-mail: hecht@physik.uni-wuerzburg.de (B.H.); wilfried.grange@tcd.ie (W.G.).
† Trinity College Dublin.
‡ These authors equally contributed.
§ University of Basel.
∥ University of Würzburg.

10.1021/jp7114862 CCC: $40.75
Published on Web 05/14/2008

$$K(\chi) = N\delta \times \left[1 - \int_0^{\chi} \mathrm{d}n\, P(n)\right] \quad (2)$$

where N is the total number of time intervals in the data set and δ is a correction factor to account for the difference between the total number of samples and the number of samples that actually contribute to the noise distribution, i.e. all bins that do not contribute to a signal burst. Note that δ can be simply obtained by comparing the maximum amplitude of both $P(n)$ and $H(n)$. For $\chi \rightarrow \infty$, the number of false positive events $K(\chi)$ approaches zero, as expected. For a finite threshold, χ, $K(\chi)$ is different from zero but can always be made sufficiently small by choosing a larger χ. We may for example define a threshold $\hat{\chi}$ by the implicit equation

$$K(\hat{\chi}) = \alpha \quad (3)$$

which corresponds to the detection of, at most, α false positive events in the N bins of the time series.

Since this analysis depends on the precise knowledge of the background probability distribution, we may conclude that the problem of distinguishing transient events from the background is reduced to the task of finding a sufficiently good estimate for the probability distribution of the background alone. The general shape of the noise distribution function can typically be determined from a "blank" experiment in which no signal bursts occur. However, the actual parameters of the distribution function, its moments, will typically depend on various experimental parameters such as the concentration of analyte molecules in an optical single-molecule detection experiment.

To find an estimate for the noise distribution function in the given data set we propose using an iterative method. In the first iteration step, the original data set is used to calculate an estimate for the mean, μ_1, and the standard deviation, σ_1, for the true μ and σ that characterize the noise. Since μ_1 and σ_1 are calculated for the entire data set including peaks well above the noise level, we expect that μ_1 and σ_1 will overestimate the true μ and σ. Naturally, we fail in this first iteration to accurately describe the contribution of noise $P(n)$ to the histogram $H(n)$. However, we may still use μ_1 to obtain a first estimate for the noise distribution

$$P_1(n) = \frac{e^{-\mu_1}{\mu_1}^n}{n!} \quad (4)$$

which may then be used to calculate a first estimate $K_1(\chi)$ for the true $K(\chi)$.

$$K_1(\chi) = N\delta \times \left[1 - \int_0^{\chi} \mathrm{d}n\, P_1(n)\right] \quad (5)$$

Similarly, we can define the following quantity:

$$K_H(\chi) = N \times \left[1 - \int_0^{\chi} \mathrm{d}n\, H(n)\right] \quad (6)$$

which is the analog of eq 2 however using the total histogram $H(n)$ of the time trace of Figure 1b instead of the background $P(n)$ alone. Now $K_1(\chi)$ is used to calculated a first estimate $\hat{\chi}_1$ for the true threshold value, $\hat{\chi}$, by invoking the analogue to eq 3 for $K_1(\chi)$.

Once a first estimate for the threshold, $\hat{\chi}_1$, is determined, the next step consists of counting fluorescence bursts with count rates above $\hat{\chi}_1$. Identifying peaks in data sets is generally a difficult procedure which requires well-designed algorithms. Here, we use a Labview routine (Peak Detector, Labview, National Instruments) based on an algorithm that fits a quadratic polynomial to a sequence of data points above $\hat{\chi}_1$. The main inputs of the routine are the threshold $\hat{\chi}_1$ and the peak width M that controls smoothing of the data when searching for peaks. Setting a small M number allows a finer resolution of the search for transient events but is prone to the detection of multiple peaks due to fluctuations on top of broader bursts. On the contrary, a too large value of M prevents the detection of short bursts. To overcome the limitations of either situation, peak detection is performed as follows: the number of consecutive data points is gradually decreased starting from a predefined

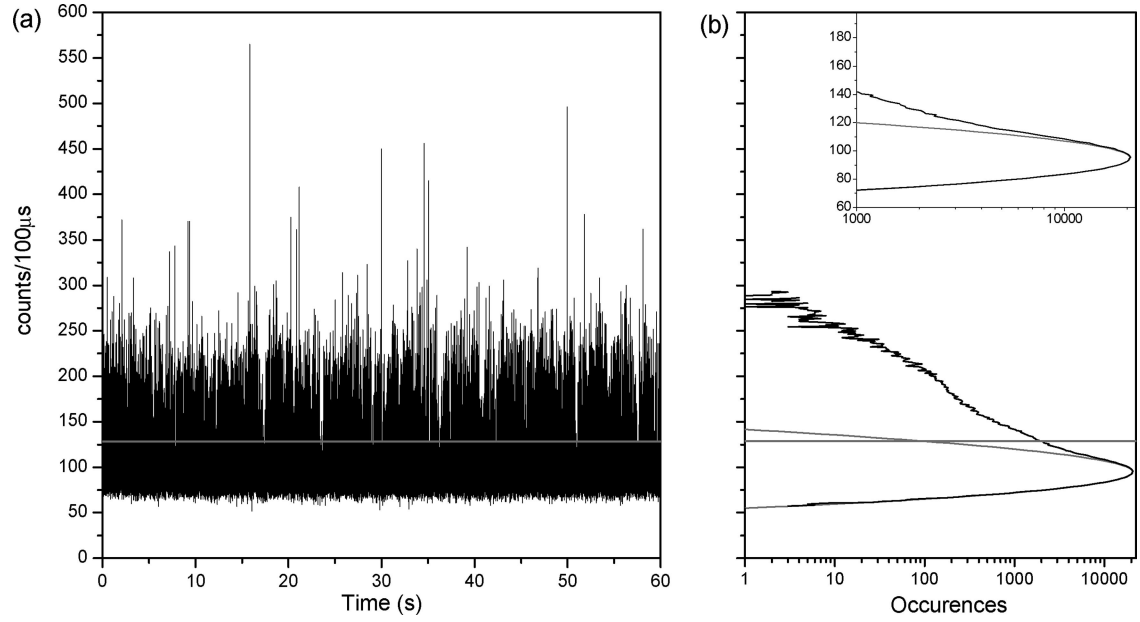

Figure 1. Time trace and histogram of a single-molecule fluorescence experiment. (a) Experimental time trace (bin width = 100 μs, $N = 6 \times 10^5$) showing fluorescence bursts on top of a strong Poissonian background. (b) Histogram of the time trace in logarithmic scale. The fluorescence bursts lead to a characteristic deviation from Poissonian statistics. The horizontal line shows the threshold level above which signals are counted as bursts. The grey curve plotted together with the histogram is the best estimate for the noise probability distribution obtained by calculating the mean of the noise after removing bursts above the optimum threshold (see text). The inset shows a zoom of the calculated histograms for both the experimental trace and the result of the algorithm.

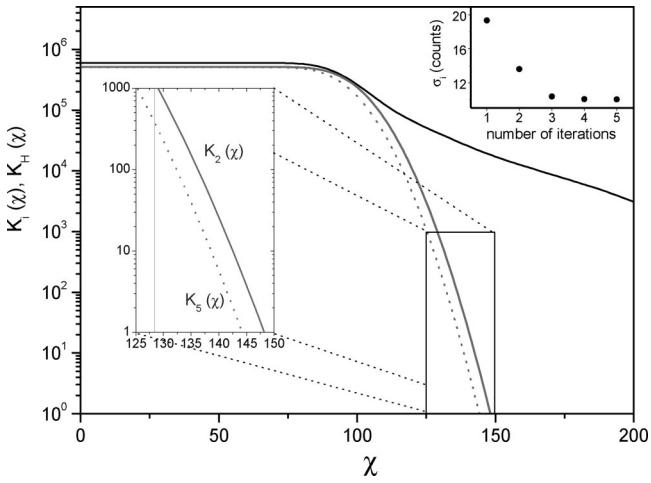

Figure 2. Visualization of the $K_i(\chi)$ and $K_H(\chi)$ (black, solid). Already, the second estimate of $K(\chi)$ (black) obtained by calculating the mean of the time trace of Figure 1 provides a good estimate for the threshold. After five iterations, all fluorescence bursts are eliminated ($K_5(\chi)$, black, dashed−dotted). The inset shows the development of σ_i for $i = 1$, 2, 3, 4, and 5. Calculations were performed for a number of false-positives of 375 and a maximum peak width of 150 (see text).

maximum M_{max}. For each value of M, the number of detected bursts is stored. The Labview routine outputs the location, amplitude, and the second derivative of the peaks but does not give any indication on the actual width of the peak, which can be significantly smaller than M. Assuming that a second order fit of the Labview routine approximates the peak, the total width, M^*, of the peak can be derived knowing both the amplitude A and second derivative S of the peak:

$$M^* = 2 \times \sqrt{\frac{2(A - \mu_i)}{S}} \tag{7}$$

Here, μ_i denotes the mean obtained at the ith iteration step. Using this procedure, we find that the actual number of detected peaks does not depend on M_{max} as long as the latter is larger than the maximum burst width that occurs in the original data set. Each detected burst is then removed from the data by removing the respective bins from the data set. After ($M_{max} - 3$) runs (3 being the minimal width of the Labview routine) of the burst detection routine, all bursts above $\hat{\chi}_1$ have been counted and removed. The remaining data set consists of the background noise plus a few bursts with amplitudes smaller than $\hat{\chi}_1$. In the second iteration step, the truncated data set obtained in the first iteration is used to calculate new estimates, μ_2 and σ_2, that better characterize the probability distribution of the noise. As a consequence, more bursts are found in this second iteration step when applying the burst finding algorithm. After i iteration steps, μ_i (σ_i) converges to a stable minimum μ (σ), which then provides a very good estimate for the parameters describing the true histogram of the background noise $P(n)$, eq 1. In practice, it is found that the algorithm converges extremely fast. As can be seen in the inset of Figure 2, the standard deviation of the truncated data set is stable already after three iterations. The resulting best estimate for the noise distribution using the parameter μ_5 is plotted in Figure 1 together with the histogram of the time trace. A remarkably good agreement is found using $\alpha = 375$ (see below).

Discussion

To investigate the reliability of the proposed algorithm, we have applied a Monte Carlo simulation of the burst counting process. To this end, we generate artificial time traces (6×10^5 bins) consisting of a Poissonian noise ($\mu = 60$, $\sigma = \sqrt{\mu} \sim 7.75$) with superimposed bursts. The height distribution of the bursts consists of a Gaussian distribution with variable amplitudes (i.e. the number of bursts is varied from 10 to 2250) and means (30, 40, 50) and a constant σ value of 10. The generated bursts are added to the time trace at random times. In addition, we use an artificial trace consisting of a Poissonian noise ($\mu = 60$) plus bursts of height 120 added to the noise to test the peak recovery capabilities of the algorithm. For each generated trace, the burst detection algorithm is applied using different α values (i.e., a different number of tolerable false positive events). Figure 3a shows the result of such simulations using α values of 1 and 33, respectively. From these simulations, important conclusions can be drawn. First, we observe that the number of detected bursts deviates from a purely linear behavior when the number of generated bursts is larger than ∼1500. This behavior (also observed for a fixed peak height distribution, bottom triangle) is not a limitation of the present algorithm but results from the fact that at high burst densities there is a significant probability of neighboring bursts overlapping and being then counted as single events. Second, the simulation shows that the number of detected bursts is smaller than the number of generated bursts and that this effect is less pronounced when α is increased. This behavior has different origins. When the mean of the burst distribution is close to that of the noise distribution, it has to be expected that the finite overlap between the two distributions tends to decrease significantly the number of detectable and detected bursts. More information can be gained by plotting the σ value (as reported by the simulation) as a function of α. As seen in Figure 3b and for a large number of generated bursts, σ deviates strongly from the expected value of 7.75. This simply means that, at low α values and for bursts distributions that overlap significantly with the noise distribution, we fail to reproduce quantitatively the latter. Considering 250 generated bursts (upper panel) and a Gaussian distribution with a mean of 30, we see for example that $\sigma = 7.84$ at $\alpha = 1$. This relatively bad agreement results in a low number of detected bursts (Figure 3a, upper panel).

With increasing α, σ approaches the expected value of 7.75 and therefore yields a higher number of peaks to be recovered. Again, these findings are not a limitation of the algorithm because we are always able to find an α value (α*) that correctly reproduces the noise distribution. In a real experiment, the noise distribution is unknown but α* can always be determined by comparing $H(n)$ and $P(n)$. For this purpose, we calculate the root mean squared error (RMSE) that is a well-accepted parameter to estimate the goodness of a fit. Because $P(n)$ can strongly deviate from $H(n)$ at large n values, the RMSE coefficient is calculated up to μ. Let us point out that the RMSE calculation could lead to large errors when the histogram $H(n)$ is poorly defined. This limitation therefore imposes the use of long enough experimental records for which the noise background is well defined. For $\alpha \ll \alpha^*$ ($\alpha \gg \alpha^*$), it is expected that the RMSE coefficient is large because of the incorrect value of σ . However, the RMSE should reach a minimum for $\alpha \sim \alpha^*$. This is exactly what is found in the simulation, where we see that the minimum of α corresponds to a σ value of 7.75 (Figures 3 and 4). Considering the experimental trace displayed in Figure 1, the optimal number of false positive events is found to be relatively large (375) due to both the high number of peaks (∼7500) and the considerable overlap of the peak distribution with the noise. The reliability of the detection method can be estimated by calculating the ratio α* to P^* (the number of peaks

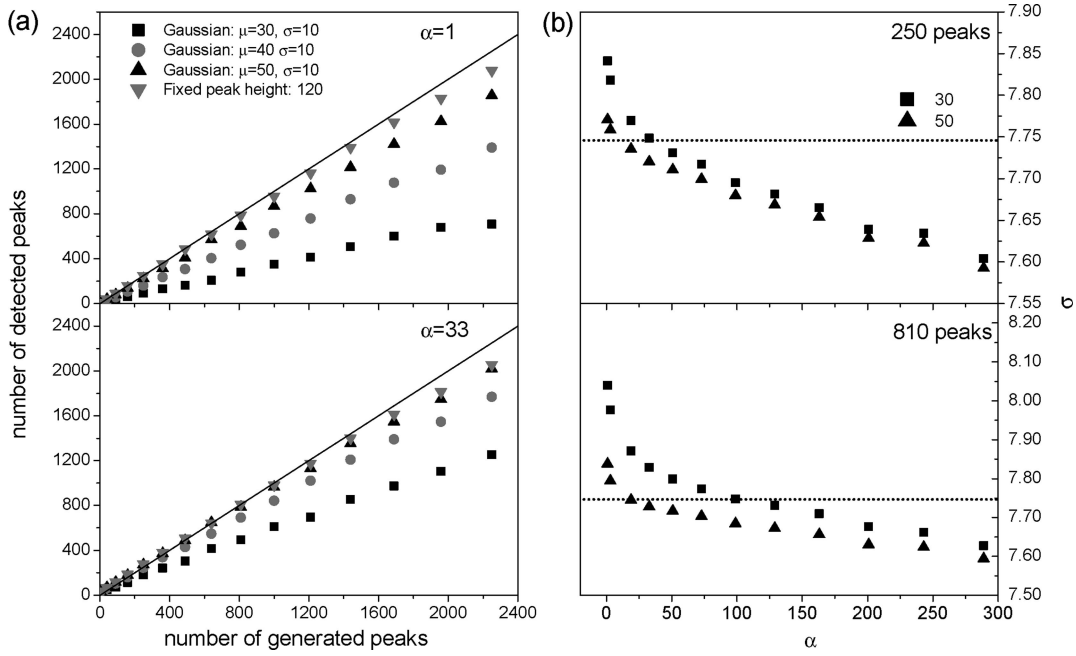

Figure 3. Monte Carlo simulations of the burst detection process. (a) Number of bursts recovered by applying the burst detection algorithm plotted against the number of artificially generated bursts (from 10 to 2250). The distribution of the generated bursts is Gaussian (with means of 30, 40, and 50 (squares, circles, and upper triangles) and a standard deviation of 10) and is added to a Poissonian noise with a mean of 60. Also shown is the result for a distribution that consists of a fixed peak height of 120 (bottom triangles) added to a Poissonian noise with a mean of 60. The simulations are performed for a number of false-positive events α of 1 and 33 (upper and lower panel, respectively). The number of detected peaks increases with α, yielding a better estimate of the number of bursts in the distribution (see text). (b) Standard deviation plotted as a function of the number of false-positive events α for a Gaussian burst distribution with mean 30 and 50, respectively: (upper panel) 250 generated peaks, (lower panel) 810 generated peaks. For a large number of generated peaks and bursts distributions that considerably overlap with that of the noise, the algorithm fails to correctly describe the noise distribution (expected σ of $\sqrt{60} \sim 7.75$), resulting in a low number of detected peaks for $\alpha = 1$.

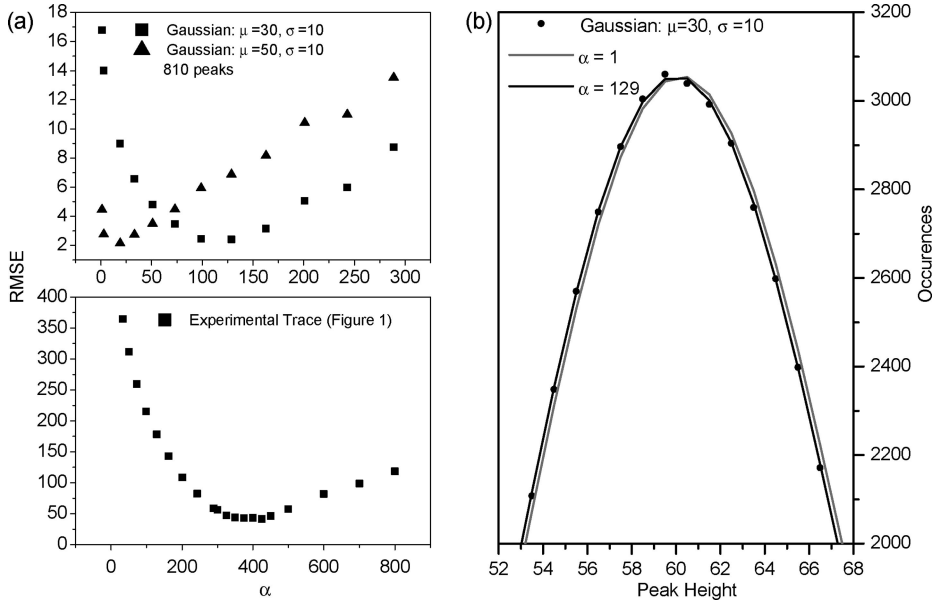

Figure 4. Goodness of the fit of the noise distribution. (a) Root-mean-square error (RMSE) as a function of the number of false positive events calculated for both the simulations (upper panel; see Figure 3) and the experimental trace of Figure 1 (lower panel). For the simulations, the RMSE reaches a minimum for a σ value of ~7.75 (the value used in the simulation for the noise distribution). (b) Comparison between the histogram of the noise (as calculated with the algorithm for different number of false positive events α) and the histogram of the artificial trace used in the Monte Carlo simulation (circles, Gaussian with $\mu = 30$, $\sigma = 10$, number of peaks 810). For $\alpha = 129$, a perfect agreement is found.

recovered at this particular α value). For the experimental trace analyzed here, we find a value of 0.05, which is reasonably good. Of course, an increase in the number of peaks or a decrease in the mean of the peak distribution would lead to lower degree of confidence. For instance, the ratio α* to P^* is about 0.2 (129/639) for a peak distribution with a mean of 30, a standard deviation of 10, and a number of generated peaks of 810. For this particular α value of 129 (Figure 3), the algorithm leads however to an almost perfect representation of the noise distribution (Figure 4).

As mentioned previously, the method is applicable to any noise distribution as long as its shape is known (Poissonian, Gaussian, . . .). For practical use, the method has to be used as follows: the number of peaks P is found from the iterative

procedure presented above (allowing the determination of the moments of the noise distribution). This calculation should be repeated for different α values (wrong-positives). Finally, the goodness of the fit (the RMSE coefficient calculated up to the mean of the noise distribution) can be plotted against α and a minimum for α can be precisely determined.

Conclusion

We have introduced an algorithm that is able to faithfully recover transient events in the presence of significant stationary noise. The method is based on the determination of an optimal detection threshold that avoids the detection of false positive events while recovering as many of the signal bursts as possible. We have demonstrated that the proposed algorithm detection allows counting of single-molecule fluorescence bursts in presence of a strong background noise. It is important to stress that the algorithm relies on robust statistical assumptions, allowing the counting of peaks with a minimum and quantifiable error. In the algorithm presented in this paper, no assumption is made on the shape or height of the burst-height distribution. Moreover, the only free parameter of the algorithm is the initial width *M*, that can be determined by finding the maximum width of the peaks in the trace. However, this can also be avoided by using a large value for *M* at the expense of larger computing times.

Acknowledgment. The authors acknowledge Profs. H.-J. Güntherodt and J. Flammer for continuous support. Financial support by the Swiss National Science Foundation via the National Center of Competence in Research in Nanoscale Science (NCCR) and a research professorship for one of the authors (B.H.) is greatfully acknowledged. Further financial support by the Wolfermann-Nägeli-Stiftung and the Centre for Research on Adaptive Nanostructures and Nanodevices (CRANN) is greatfully acknowledged. W.G. thanks Ronan Daly (CRANN) for critical reading of the manuscript.

References and Notes

(1) Zander, C.; Enderlein, J.; Keller, R. A. *Single-Molecule Detection in Solution: Methods and Applications* Wiley-VCH,: Berlin, 2002.
(2) Basché, T.; Moerner, W. E.; Orrit, M.; Wild, U. P. *Single Molecule Optical Detection, Imaging, and Spectroscopy* (VCH, Weinheim, 1997).
(3) Soper, S. A.; Mattingly, Q. L.; Vegunta, P. *Anal. Chem.* **1993**, *65* (6), 740–747.
(4) Rissin, D. M.; Walt, D. R. *Nano Lett.* **2006**, *6* (3), 520–523.
(5) Irizarry, R. A.; Hobbs, B.; Collin, F.; Beazer-Barclay, Y. D.; Antonellis, K. J.; Scherf, U.; Speed, T. P. *Biostatistics* **2003**, *4*, 249–264.
(6) Zhang, Y.; Bahns, J. T.; Jin, Q.; Divan, R.; Chen, L. *Anal. Biochem.* **2006**, *356* (2), 161–170.
(7) Ignatovich, F. V.; Novotny, L. *Phys. Rev. Lett.* **2006**, *96*, 013901.
(8) Collinson, M. M.; Wightman, R. M. *Science* **1995**, *268*, 1883–1885.
(9) Fan, F. R. F.; Bard, A. J. *Science* **1995**, *267*, 871–874.

JP7114862