BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 2 2012, pages 296–297
doi:10.1093/bioinformatics/btr632

Data and text mining

Advance Access publication November 15, 2011

NoRSE: noise reduction and state evaluator for high-frequency single event traces

Nigel F. Reuel[1], Peter Bojo[1], Jingqing Zhang[1], Ardemis A. Boghossian[1], Jin-Ho Ahn[1], Jong-Ho Kim[2] and Michael S. Strano[1,*]

[1]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA and [2]Department of Chemical Engineering, Hanyang University, Ansan 426-791, Republic of Korea

Associate Editor: John Quackenbush

ABSTRACT

Summary: NoRSE was developed to analyze high-frequency datasets collected from multistate, dynamic experiments, such as molecular adsorption and desorption onto carbon nanotubes. As technology improves sampling frequency, these stochastic datasets become increasingly large with faster dynamic events. More efficient algorithms are needed to accurately locate the unique states in each time trace. NoRSE adapts and optimizes a previously published noise reduction algorithm and uses a custom peak flagging routine to rapidly identify unique event states. The algorithm is explained using experimental data from our lab and its fitting accuracy and efficiency are then shown with a generalized model of stochastic datasets. The algorithm is compared to another recently published state finding algorithm and is found to be 27 times faster and more accurate over 55% of the generalized experimental space. NoRSE is written as an M-file for Matlab.

Availability: http://web.mit.edu/stranogroup/NoRSE.txt

Contact: strano@mit.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on September 13, 2011; revised on November 7, 2011; accepted on November 9, 2011

1 INTRODUCTION

As stochastic biological studies are preformed at increasingly smaller length and time scales, the analysis of large, noisy datasets is becoming an increasingly common problem. In our laboratory, these sets consist of time traces gathered from monitoring the fluorescence of single-walled carbon nanotube (SWNT)-based sensors. As target molecules bind to a single sensor, the fluorescence is quenched in a stepwise manner (Cognet *et al*., 2007), resulting in unique fluorescent states that correspond to the number of bound sites along the length of the nanotube (Jin *et al*., 2008). Although NoRSE was created with our fluorescence datasets in mind, the algorithm can be readily applied to other, analogous biological experiments that contain event states. This is demonstrated by an error analysis of generalized stochastic traces as shown below.

*To whom correspondence should be addressed.

2 ALGORITHM EXPLANATION

In this note, data from our SWNT-based, protein–protein sensors are analyzed. Excited SWNT fluoresce due to their unique band gap structure (Bachilo *et al.*, 2002) and can be quenched by specific molecules or chelating groups (Jin *et al*., 2008). At a single-molecule level, the quenching is exhibited in a stepwise manner as each of the limited, exciton excursion distances of the SWNT are occupied (Cognet *et al.*, 2007). Our 900 nm sensor exhibits a maximum of 10 binding events or 10 step levels. The fluorescence of each SWNT is monitored for 3000 time steps and NoRSE efficiently resolves the bound states. NoRSE imports the data traces $X_1(t)$, $X_2(t)$, … $X_n(t)$ and normalizes them by the maximum intensity such that $\{X_i : 0 < X_i < 1\}$ (Fig. 1A—top trace is a control and the bottom trace exhibits binding events).

$X_i(t)$ is then noise reduced (Fig. 1B) by the Chung and Kennedy algorithm (Chung *et al.*, 1991). Their forward–backward, non-linear filtering technique was designed to preserve high-frequency step events surrounded by background noise. In brief, the algorithm searches forward and backwards from each data point using a bank of window sizes to weigh the probability of a sharp step occurring. Their algorithm contains three parameters (p- which sets the sharpness of state transitions, N- which specifies a set of forward and backward windows and M- which sets how much data the N windows are run against), which were optimized in the original work for patch-clamp experiments (Chung *et al.*, 1991) and subsequently for protein folding experiments (Haran, 2004). We used Monte Carlo simulations of generic, noisy, high-frequency event traces to further explore this parameter optimization (Supplementary Material A). Generally, we found that an increasing value of p and N and a decreasing value of M improved the noise reduction algorithm's ability to reconstruct the original trace, but the improvement was marginal. Thus, the program is set to run with p, N and M values of 40, [4 8 16 32], and 10, respectively, for all types of experimental traces.

After noise reduction of $X_i(t)$, all point histograms with 200 bins are generated for $X_i(t)$ (Fig. 1C). The distinct groups of histogram peaks represent event states but vary in height and width depending on their frequency and uniqueness, respectively. A detailed explanation of this peak flagging routine is given in the Supplementary Material, but briefly all potential peaks are flagged and then logic is used to determine which peaks are unique and which should be combined with neighboring peaks. The result is a vector of significant peak bin numbers, which are then translated back to the corresponding signal levels of $X_i(t)$. These are the unique

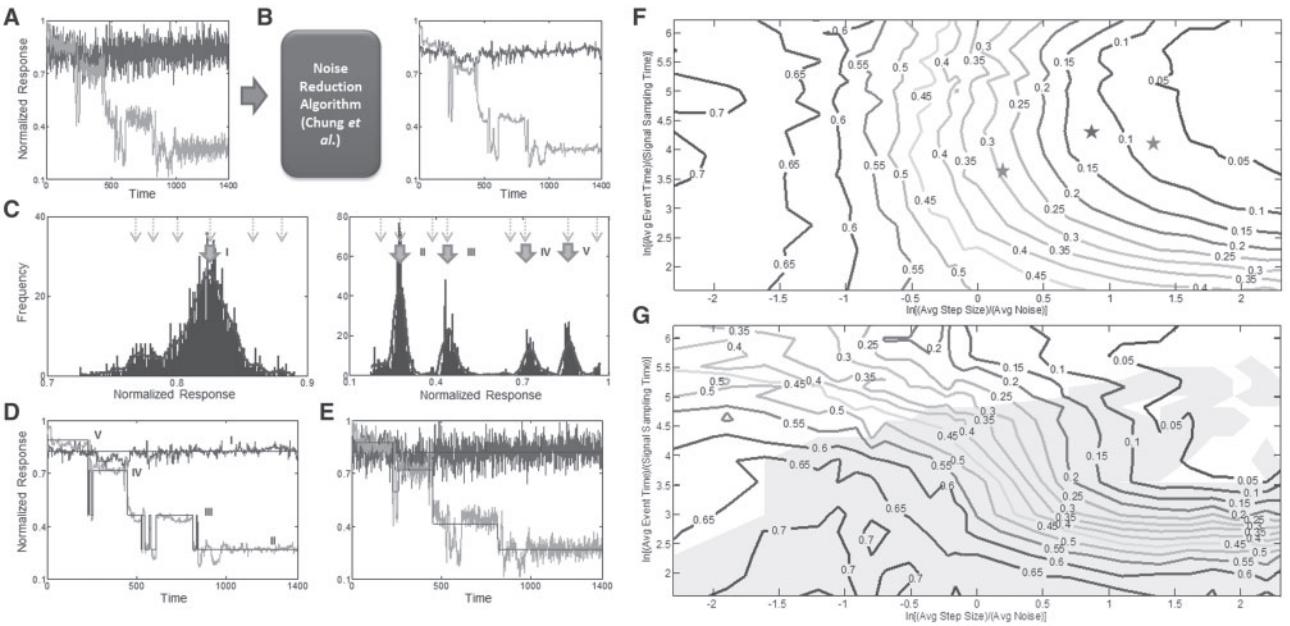

Fig. 1. Explanation and evaluation of NoRSE. (**A**) Control (blue) and signal (green) traces are normalized and (**B**) noise reduced by Chung *et al.* algorithm. (**C**) All point histogram peaks of each trace are identified as unique event states to which the trace is fit (**D**). (**E**) Corresponding SFA fit of these traces. (**F**) The fit error percentage of NoRSE over a general experimental space (green star = general region for patch clamp experiments, red = FRET protein folding and orange = SWNT protein–protein sensors). (**G**) Error percentage of SFA algorithm. Shaded area is the experimental space in which NoRSE outperforms SFA.

state levels for the trace. Finally, each time point of the noise reduced trace of $X_i(t)$ is compared with the possible state levels and the one that fits closest is assigned. Qualitatively, the fit looks good for our experimental data (Fig. 1D), but we were interested in a quantitative method of determining the error as well as finding its applicability to other types of data that contain step events. This analysis is done in the following generalized model.

3 ERROR ANALYSIS OF GENERALIZED MODEL

Stochastic traces that contain unique event states can be generalized with three basic parameters: A = (tendency upwards movement / tendency downwards movement), B = (average event time / signal sampling time) and C = (average event transition size / average noise level). Further discussion on the meanings of these parameters is provided in the Supplementary Material, but many types of single-event experiments can be cast on this generalized experimental space (Fig. 1E). As the parameter A has less of an effect on the fitting-fidelity, it was held at the equilibrium level of 1, while the other parameters B and C were varied over the ranges of 5–500 and 0.1–10, respectively. For each parameter combination, 100 traces were generated at 1000 time steps each. The noisy trace was sent to NoRSE and the resulting fit was compared with the known generated trace. Each time step was analyzed to see if it was within 2% of the real value, if not, it was marked as an error. Finally, the average error percent for each B–C combination was reported (Fig. 1F). The contour plot maps the specific regions where the NoRSE algorithm is most accurate; it begins to have errors >10% at B values <20 and C values <2.7.

To compare NoRSE's computational efficiency and accuracy, an identical error analysis was performed on a popular 'State-Fitting Algorithm' (SFA) used for similar noisy, dynamic traces (Kerssemakers *et al.*, 2006). SFA uses a series of chi-squared best fit analyses as the number of steps (S) is iteratively increased and compared to corresponding 'anti-fits'. The resulting error contour map (Fig. 1G) indicates regions in which NoRSE outperforms SFA (>55% of the modeled space—shaded region in Fig. 1G). The physical run time of NoRSE on the 39 000+ model traces was 27 times faster. By scaling the algorithms and taking the approximate ratio of SFA to NoRSE (Supplementary Material D), we find that NoRSE is on the order of $[S^2/(K(N+3))]$ more efficient than SFA, where S is the number of SFA fit steps and K and N are parameters of the noise reduction algorithm. The two algorithms were performed on the protein–protein traces and the average run time was 25 times faster for NoRSE (32 predicted from approximate scaling).

Funding: NSF Graduate Fellowship to N.F.R.

Conflict of Interest: none declared.

REFERENCES

Bachilo,S.M. *et al.* (2002) Structure-assigned optical spectra of single-walled carbon nanotubes. *Science*, **298**, 2361–2366.

Chung,S.H. and Kennedy,R.A. (1991) Forward-backward non-linear filtering technique for extracting small biological signals from noise. *J. Neurosci. Meth.*, **40**, 71–86.

Cognet,L. *et al.* (2007) Stepwise quenching of exciton fluorescence in carbon nanotubes by single-molecule reactions. *Science*, **316**, 1465–1468.

Haran,G. (2004) Noise reduction in single-molecule fluorescence trajectories of folding proteins. *Chem. Phys.*, **307**, 137–145.

Jin,H. *et al.* (2008) Stochastic analysis of stepwise fluorescence quenching reactions on single-walled carbon nanotubes: single molecule sensors. *Nano Lett.*, **8**, 4299–4304.

Kerssemakers,J.W.J. *et al.* (2006) Assembly dynamics of microtubules at molecular resolution. *Nature*, **442**, 709–712.