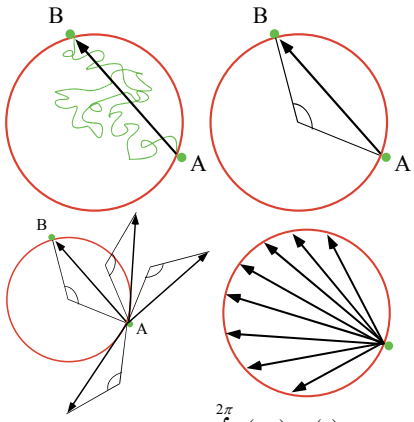
Chemometric Approaches to Quantifying Single-Molecule Surface Diffusion with the Single Event Duration Histogram

Michael J. Culbertson, Daniel L. Burden

Wheaton College, Chemistry Department, Wheaton, IL 60187

Abstract: A widely heralded feature of **single-molecule measurements** is the ability to characterize rare and disparate behaviors that are normally lost to ensemble averaging. **Confocal microscopy**, which measures the fluorescence of tagged molecules moving in and out of a focused laser, has particular potential to extract these features due to its high optical throughput and time resolution. However, Fluorescence Correlation Spectroscopy (FCS), the primary method of analyzing surface diffusion information collected with this approach, is limited under **heterogeneous conditions** because it is inherently an ensemble-averaging operation. Thus, we propose a new method for characterizing surface diffusion via Single Event Duration Histogram (SEDH). In order for the approach to be quantitatively useful, a mathematical model of the SEDH must be developed. We have created a **computer simulation** of two-dimensional random molecular motion to generate SEDH curves with varying input parameters, such as **diffusion constant**, **fluorescence intensity**, and **surface density**. In this presentation we summarize our efforts to model the histogram's response through analytical geometry, multivariate analysis, and guess-and-check prediction. Other potential methods of analysis including neural networks, genetic algorithms, non-linear multivariate analysis, and "smart" merit functions are also discussed.



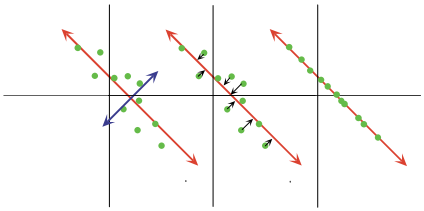
$$\mathrm{SEDH}_D = \int_0^{2\pi} A(\theta,t)\cdot\Phi(\theta)\,d\theta$$

Analytical Geometry

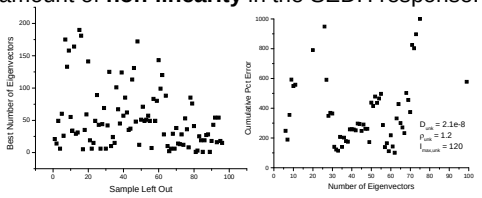
The random motion of a particle through the laser spot can be approximated by an **Arrival Time Distribution**, i.e. the probability that the particle arrives at a given distance in a given amount of time. Then, the direction of the particle's motion can be constrained using a **Flux Probability Distribution**, i.e. the probability that the particle exits with a given angular displacement.

Following this logic, an SEDH can be calculated by integrating the ATD and the FPD over all angles. Problems with this method arise from the **complex boundary conditions** imposed by the Gaussian intensity profile of the laser spot.

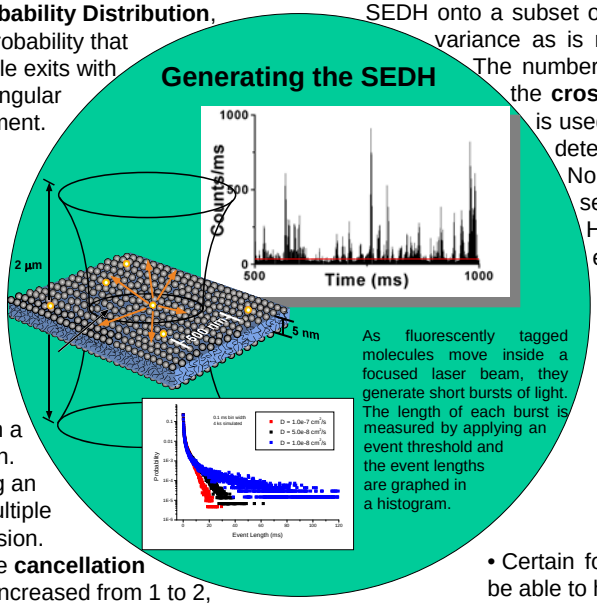
Multivariate Analysis



If the data points of the SEDH vary linearly with the input parameters (diffusion constant, surface density, fluorescence intensity), the SEDH can be interpreted using **Multiple Linear Regression**. Since the system is over-determined by ~1000 points comprising an SEDH, the data are first subjected to **Principal Component Analysis** to reduce the number of variables. PCA involves projecting the SEDH onto a subset of its **eigenvectors** which capture just as much variance as is necessary to adequately describe the system. The number of eigenvectors to use can be determined by the **cross-validation technique** in which a training set is used to create a model and a testing set is used to determine the predictive accuracy of the model. Normally, eigenvectors are gradually added to the set until the predictive error begins to decrease. However, for the SEDH, the optimal number of eigenvectors varied widely depending on the input parameters, possibly indicating a critical amount of **non-linearity** in the SEDH response.



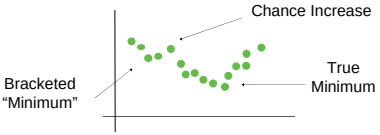
Generating the SEDH



As fluorescently tagged molecules move inside a focused laser beam, they generate short bursts of light. The length of each burst is measured by applying an event threshold and the event lengths are graphed in a histogram.

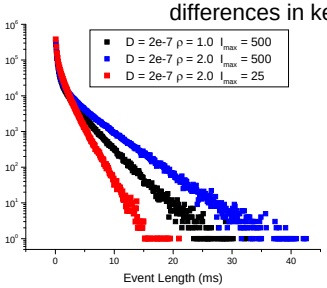
Guess-and-Check Prediction

A rather inelegant but functional method of prediction involves the same processes as in **curve fitting**. An experimental SEDH is compared with a "known" curve generated by computer simulation. The figure of comparison is then minimized using an algorithm such as the **Simplex Algorithm** for multiple dimensions or **Brent's Algorithm** for one dimension. Problems arise from undesired minima due to the **cancellation** of different parameters effects: When density is increased from 1 to 2, the curve shifts to the right, but when the maximum intensity is then reduced to from 500 to 25, the curve shifts back to the left. At some point, different sets of parameters "**cross-over**" and are indistinguishable by the merit function. Further difficulties stem from the **noise** in the merit function caused by the **randomness** of the simulation. Algorithms such as Brent's require a smooth, continuous function, but chance increases in the merit function's value can cause the algorithm to get caught up in a false minimum. Furthermore, the function is very **computationally intensive**: twenty computers work for 3 minutes to generate one SEDH from simulation.



Where next?

- Certain forms of **non-linear multivariate analysis** may be able to handle the complex behavior of the SEDH.
- New **figures of merit** may be able to identify important differences in key regions of the SEDH curve.



- An innovative **Grid Optimization Algorithm** has been developed specifically for noisy, expensive-to-evaluate functions.
- Other techniques such as **Neural Networks** and **Genetic Algorithms** may be able to detect the complex patterns in the interactions of the various input parameters.