Bayesian Analysis of Stochastic Nanopore Data

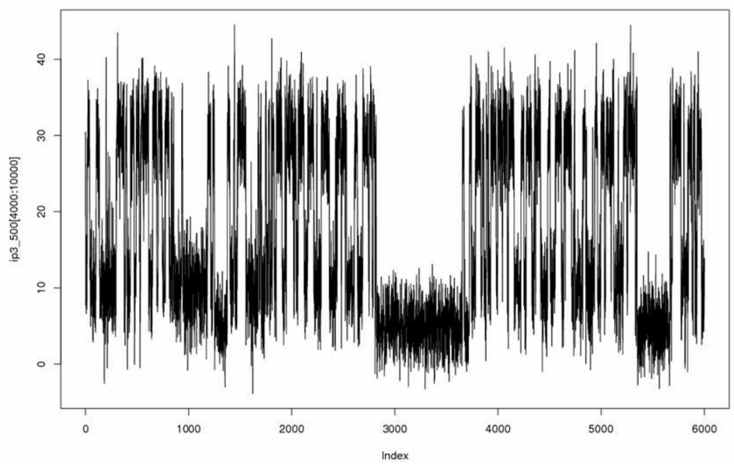
SAND2007-1509P

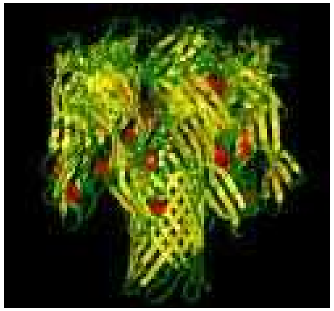
- Bayesian detection of Co-Zn mixtures concentration, based on QUB sim.
  - with quantified probability distributions
- Studies of detection probability distribution dependence on # of events and noise level
- Detector ROC curves generated from Bayesian analysis of simulated nanopore data with Co-Zn mixtures

Assessment and Challenges

- Key Challenges:
  - Dealing with signal noise, detector drift
  - Training for mixtures concentration detection
    - In the absence of sufficient empirical data
- Objectives:
  - Demonstrate detection from noisy data
  - Demonstrate detection with intermediate signal levels
  - Demonstrate single/mixed agent detection

Stochastic nanopore array data





Attributes of stochastic signal

- Frequency of transitions
- Statistics of open and closed intervals
- Current Amplitude

Challenges

- Noise of signal (SNR) and ability to identify transitions (duration and amplitude)
- Multiple states (intermediate current levels)
- Modelling stochastic behaviour of mixtures
- Response time of detector (computational cost must be low)
- Signal artifacts, baseline drift, outliers
- Trade-off Probability of Detection vs Probability of False Alarm

Applying Bayes Theorem

Identify the agents present in the sample, and estimate their concentrations. The joint posterior pdf of the unknowns is

$$p(M \mid data) = \frac{p(data \mid M) \cdot p(M)}{p(data)}$$

$$M = \{Agent(s)\, ID,\ Concentration(s)\}$$

Specification of the Likelihood Function

Predictive probability of the data (attributes of the stochastic signa

Attributes are assumed statistically independent

$$p(data \mid M) \equiv \prod p(attribute_i \mid M)$$

Modelling

Frequency of transitions, n/N. n=number of transitions. N=sample size

$$n \sim Binomial(n \mid \theta, N)$$

Duration of open/closed intervals. Dirichlet discretization into k intervals.

$$\mathbf{n} \sim Multinomial(\mathbf{n} \mid \boldsymbol{\theta}), \qquad \mathbf{n} = \{n_1, n_2, ..., n_k\} \qquad \boldsymbol{\theta} = \{\theta_1, \theta_2, ..., \theta_k\}$$

Bayesian Estimation Of The Unknown Parameters (training)

Binomial-Beta model

$$n \sim Binomial(n \mid \theta, N)$$

$$\theta \sim Beta(\theta \mid \alpha_1, \alpha_2)$$

$$p(\theta \mid n, N) \propto Binomial(n \mid \theta, N) \cdot Beta(\theta \mid \alpha_1, \alpha_2) \sim Beta(\theta \mid n + \alpha_1, N - n + \alpha_2)$$

Multinomial-Dirichlet model (multivariate generalization of Binomial-Beta model)

$$\mathbf{n} \sim Multinomial(\mathbf{n} \mid \boldsymbol{\theta})$$

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) \sim Dirichlet(\boldsymbol{\theta} \mid \boldsymbol{\alpha})$$

$$p(\boldsymbol{\theta} \mid \mathbf{n}, M) \propto Multinomial(\mathbf{n} \mid \boldsymbol{\theta}) \cdot Dirichlet(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \sim Dirichlet(\boldsymbol{\theta} \mid \mathbf{n} + \boldsymbol{\alpha})$$

$$p(\mathbf{n}_{+1}) = \int Multinomial(\mathbf{n}_{+1} \mid \boldsymbol{\theta}) \cdot Dirichlet(\boldsymbol{\theta} \mid \mathbf{n} + \boldsymbol{\alpha})\, d\boldsymbol{\theta} = NHG(\mathbf{n}_{+1} \mid \mathbf{n}_{+1} + \mathbf{n} + \boldsymbol{\alpha})$$

NHG: Negative HyperGeometric Distribution. Analytical Form

$$NHG(\mathbf{n}_{+1} \mid \mathbf{n}_{+1} + \mathbf{n} + \boldsymbol{\alpha}) = \frac{Z(\mathbf{n}_{+1} + \mathbf{n} + \boldsymbol{\alpha})}{Z(\mathbf{n} + \boldsymbol{\alpha}) \cdot M(\mathbf{n}_{+1})} \qquad Z(\mathbf{x}) = \frac{\prod_i \Gamma(x_i)}{\Gamma\left(\sum_i x_i\right)} \qquad M(\mathbf{x}) = \frac{\prod_i x_i!}{\left(\sum_i x_i\right)!}$$

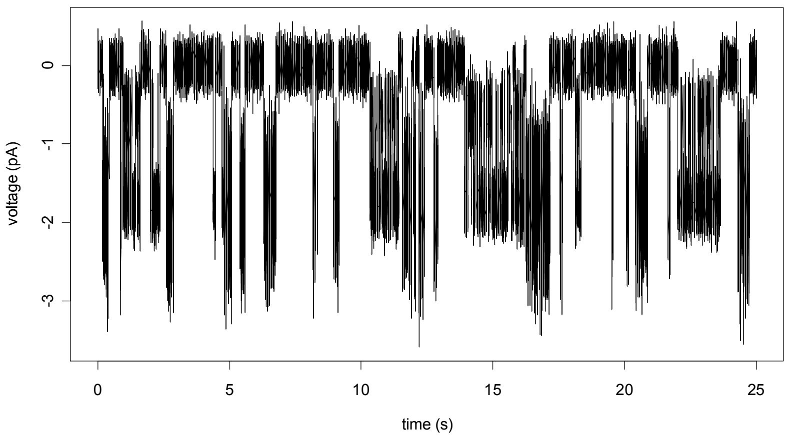
Continuous parameterization and Interpolation of counts n as a function of concentrations of mixture components M_s

$$\mathbf{n}_s = \varphi\left(M_s, \mathbf{n}_{training}, M_{training}, \xi\right)$$

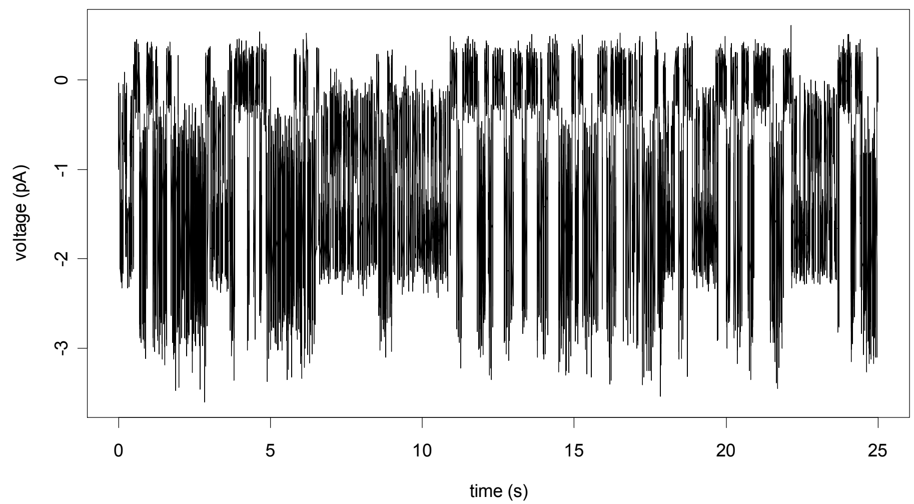
$$p(\boldsymbol{\theta} \mid \mathbf{n}_s, M_s) \approx Dirichlet\,(\mathbf{n}_s + \boldsymbol{\alpha})$$

$$p(\mathbf{n}_{+1} \mid M_s) = \int Multinomial(\mathbf{n}_{+1} \mid \boldsymbol{\theta}) \cdot Dirichlet(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) d\boldsymbol{\theta} = NHG(\mathbf{n}_{+1} \mid \mathbf{n}_{+1} + \mathbf{n}_s + \boldsymbol{\alpha})$$

Test : mixture Co-Zn, α-Hemolysin Pore
Simulated stochastic data QuB Software Suite.



Co = 2 μM, Zn = 90 nM

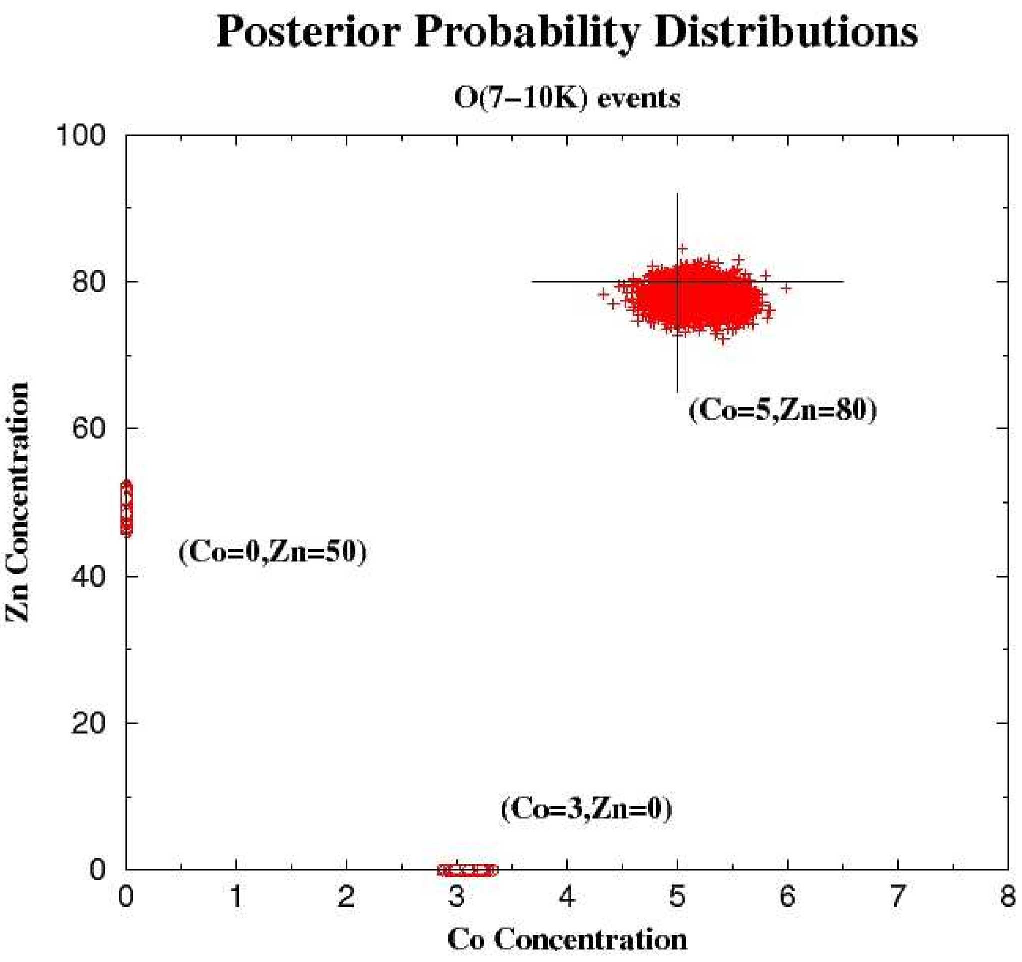


Co =8 μM, Zn= 90 nM

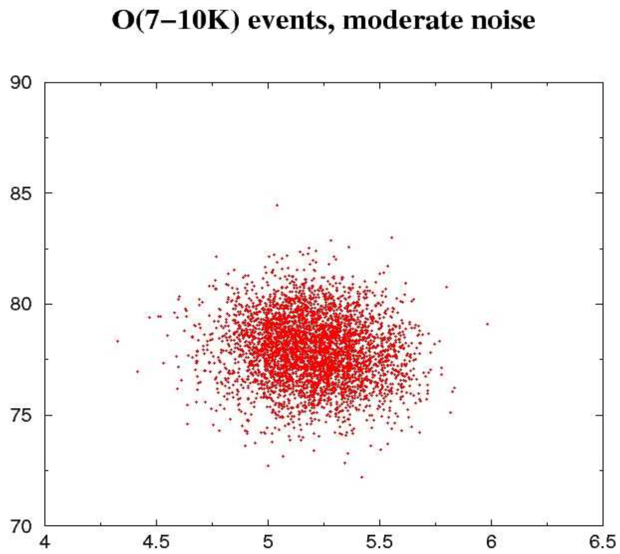
- Co: Width and frequency of top-level gaps are most useful attributes
- Zn: Width and frequency of bottom-level gaps are most useful
- Combination of top-level and bottom-level gap statistics and event frequencies is effective for mixtures of both

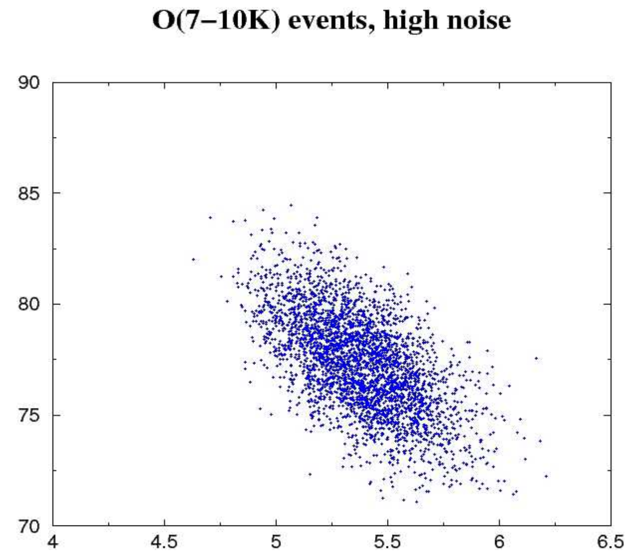
Posterior Probability Distribution in the Co:Zn Plane

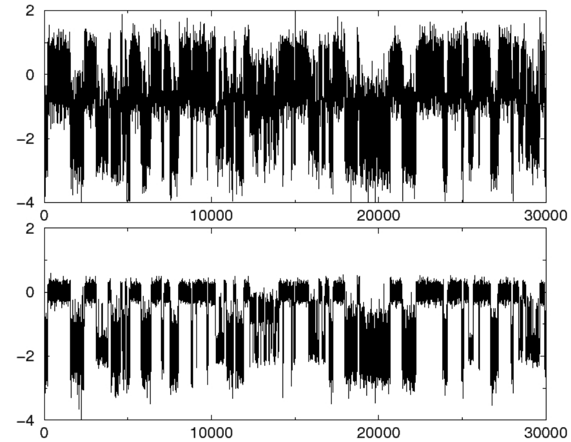
- Posterior plotted for
  - pure Co
  - pure Zn
  - Co+Zn
- Using MCMC
- P([Co],[Zn] | Data)
- Highly peaked and well centered PDFs in all cases



Effect of Noise and Number of Events on Observed Posteriors







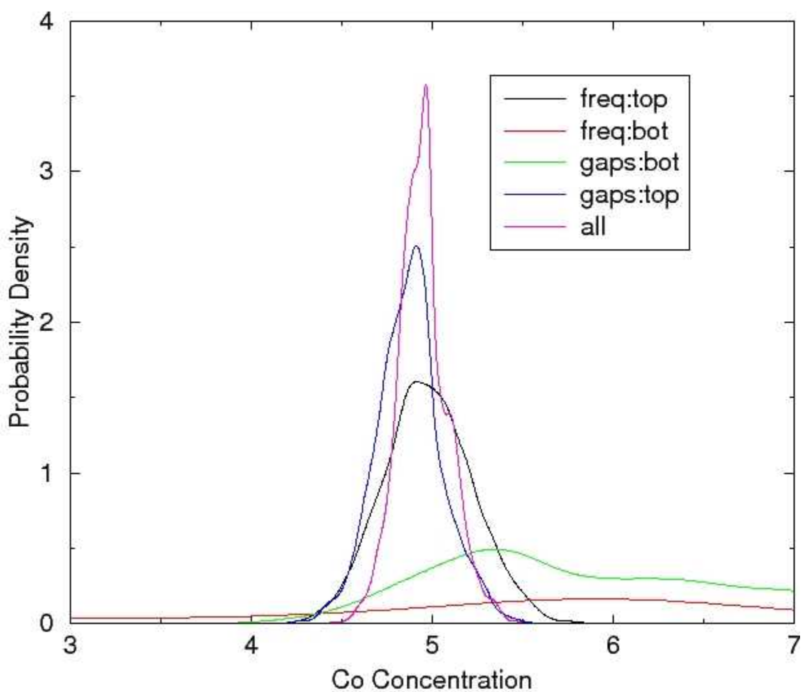
High Noise

Moderate Noise

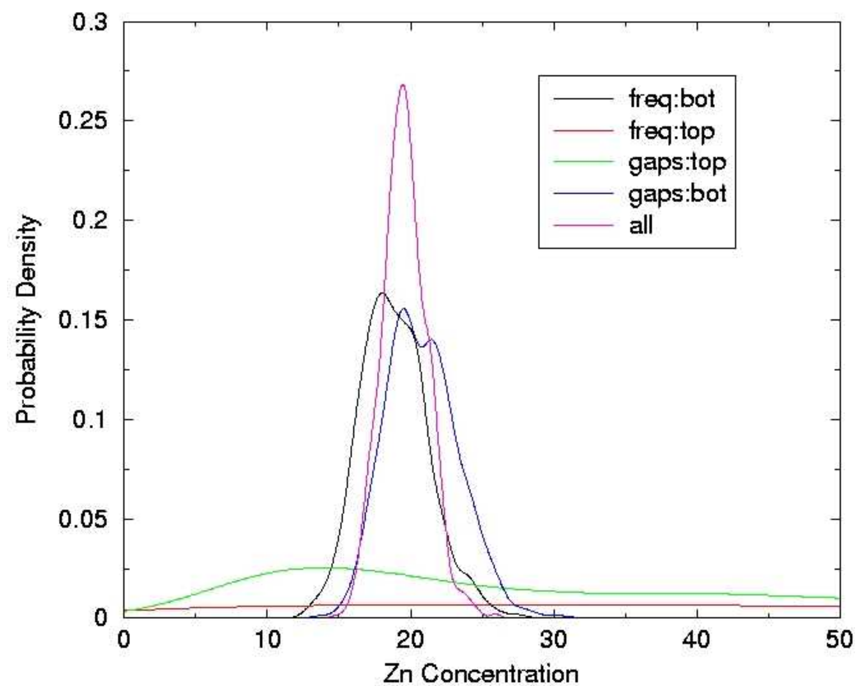
Performance of Various Attributes

• Co: Width and frequency of top-level gaps are most useful attributes
• Zn: Width and frequency of bottom-level gaps are most useful
• Combination of top-level and bottom-level gap statistics and event frequencies is effective for mixtures of both

Attributes for Co=5 Detection



Attributes for Zn=20 Detection



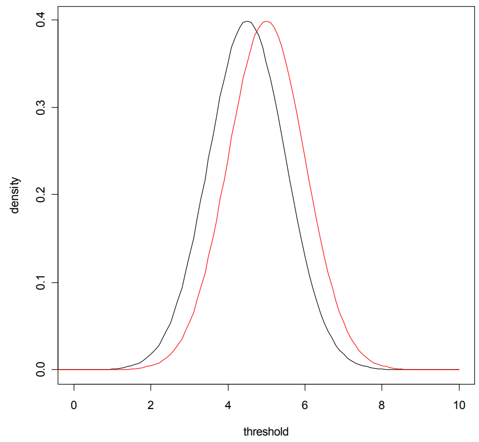
Receiver Operating Characteristic (ROC) Curves

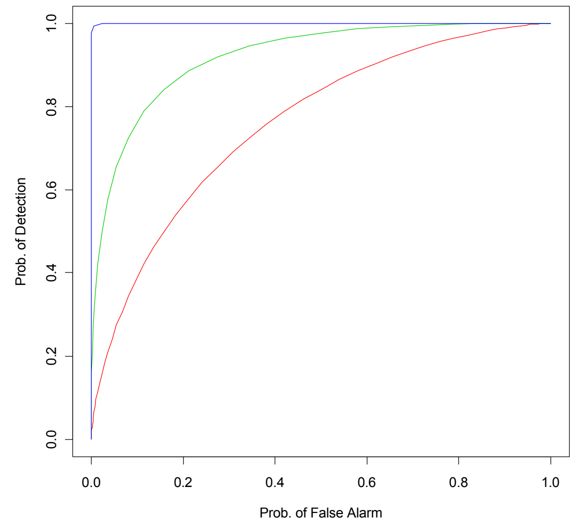
As a measure of the performance of a two-class classifier, a ROC curve shows the trade-off between sensitivity (true positives) and specificity (false negatives) for different values of a threshold.
If the threshold is the log of the Bayes Factor:

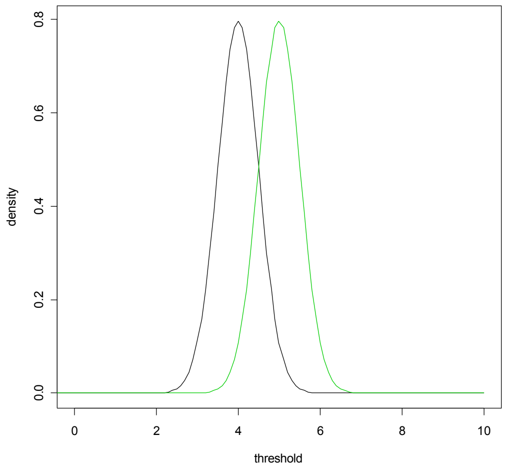
$$\log B_{YT} = \log\left(\frac{p(T \mid y_T)}{p(F \mid y_T)}\right) \qquad p(\log B_{YT}) = \int p(\log B_{YT}, y_T)\,dy_T$$

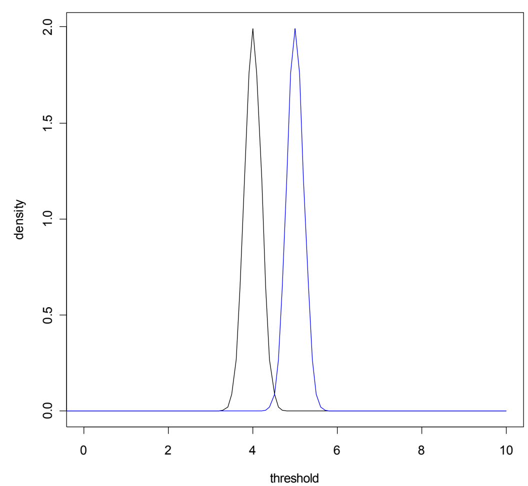
$$\log B_{YF} = \log\left(\frac{p(T \mid y_F)}{p(F \mid y_F)}\right) \qquad p(\log B_{YF}) = \int p(\log B_{YF}, y_F)\,dy_F$$

Interpretation of ROC Curves







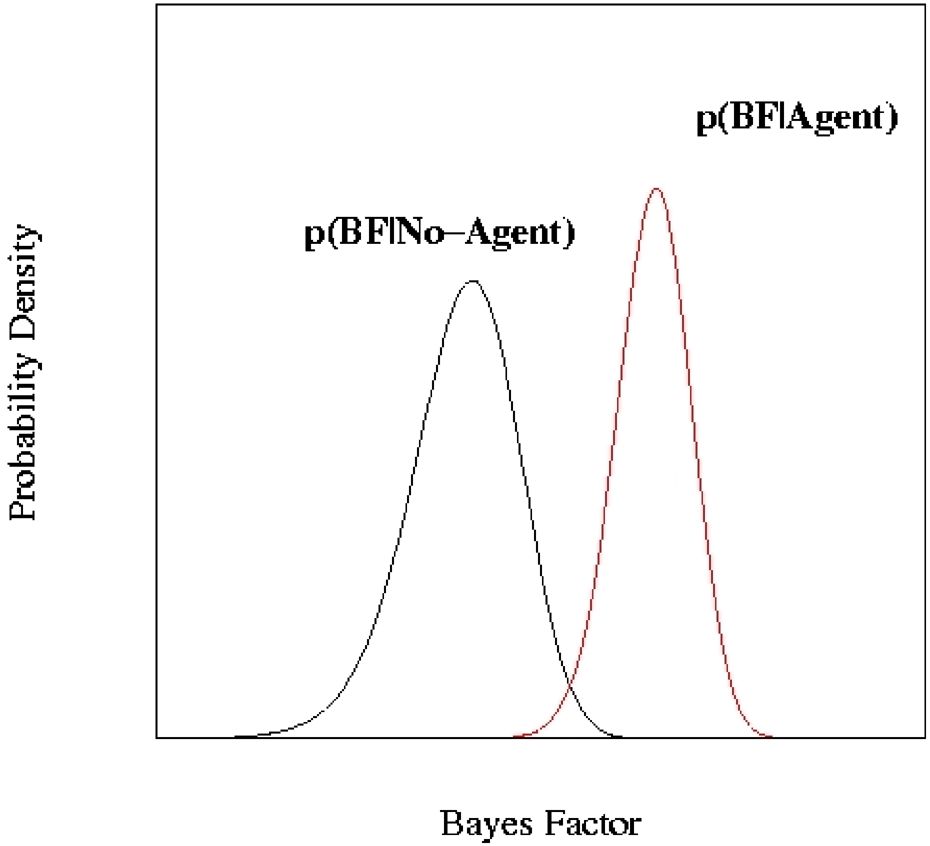


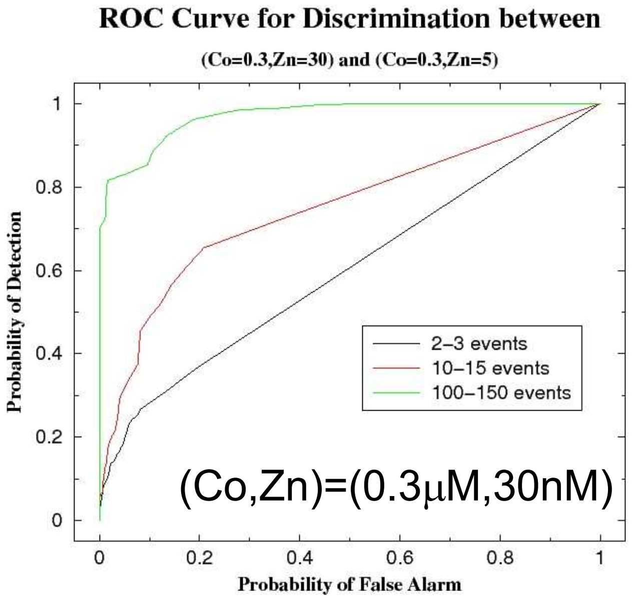
ROC Curves Construction, given Bayesian Data Analysis

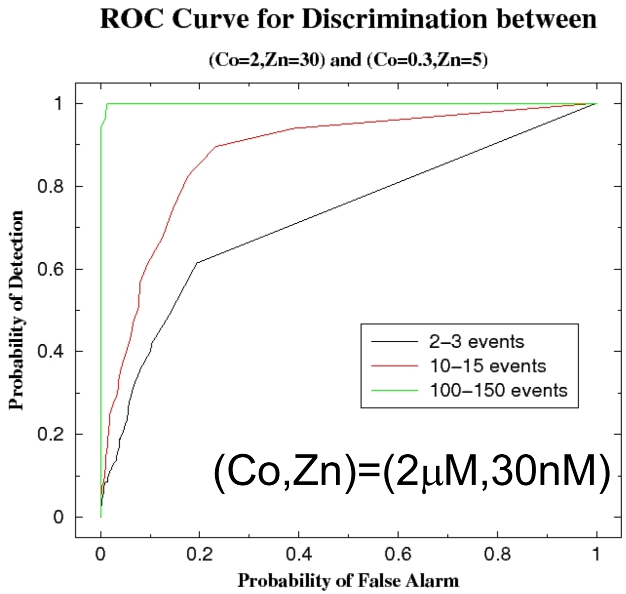
Bayes Factor (BF) defined as:

P(Agent|data)/P(No-Agent|data)

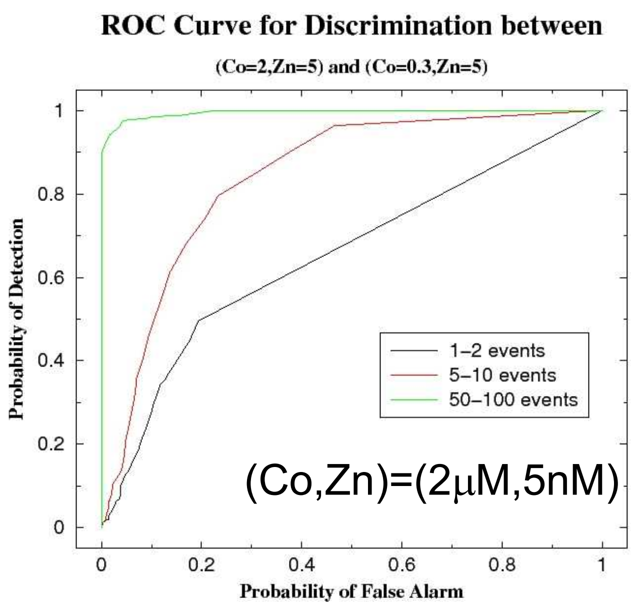
Probability distributions of BF for cases of Agent/No-Agent are used to construct ROC curves



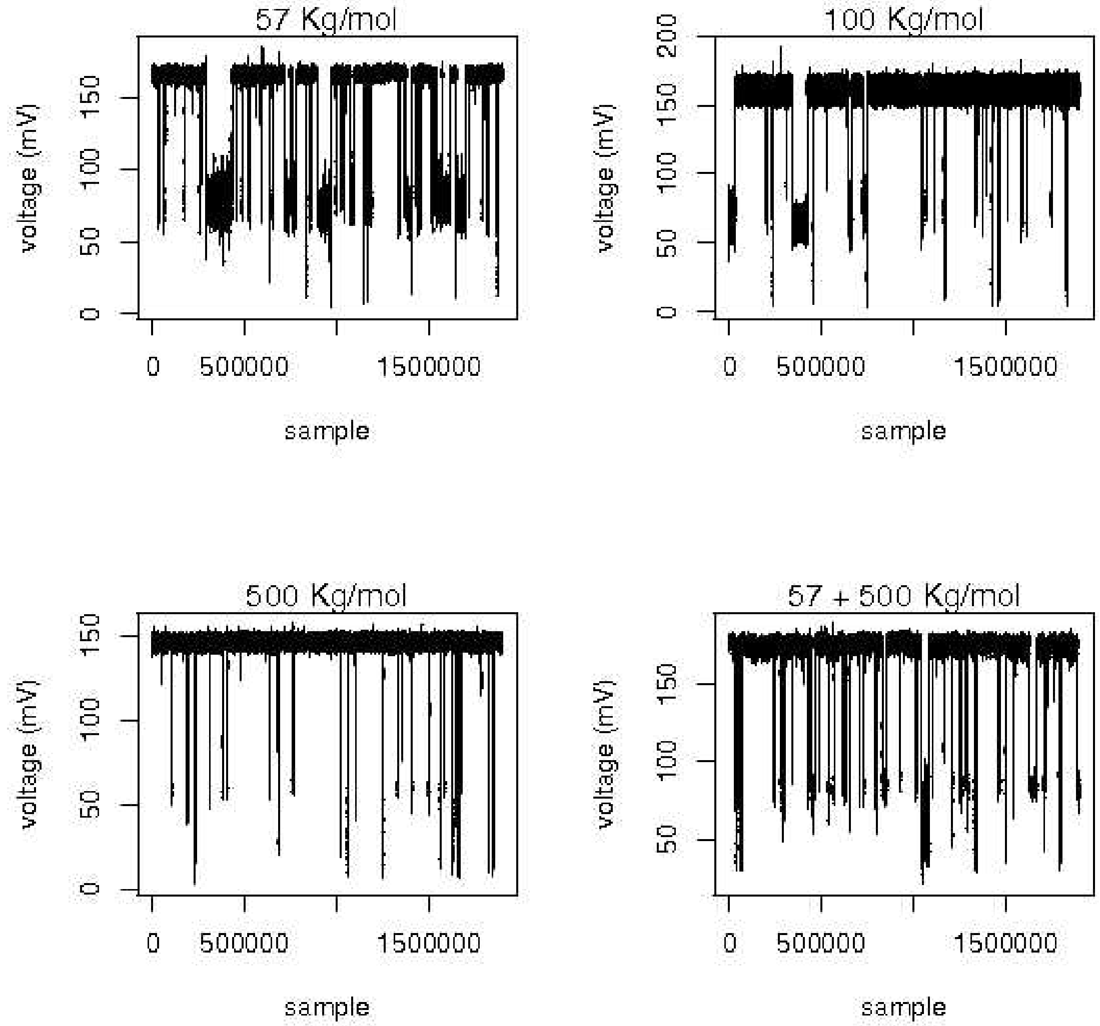




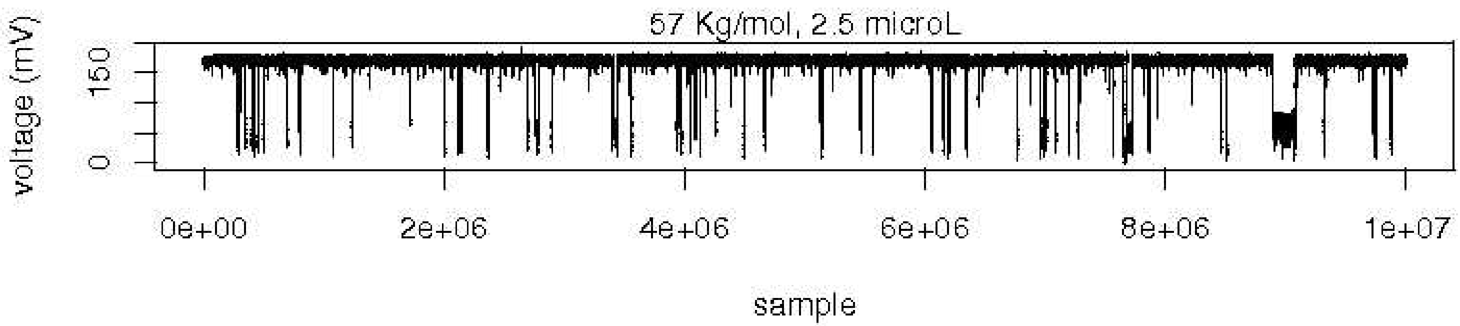
- ROC Curves constructed using Bayes Factors
  - general non-parametric PDFs
  - detection of (Co:Zn) mixtures
    - vs ~ zero levels
  - range of # of events
- O(100) events sufficient for excellent performance

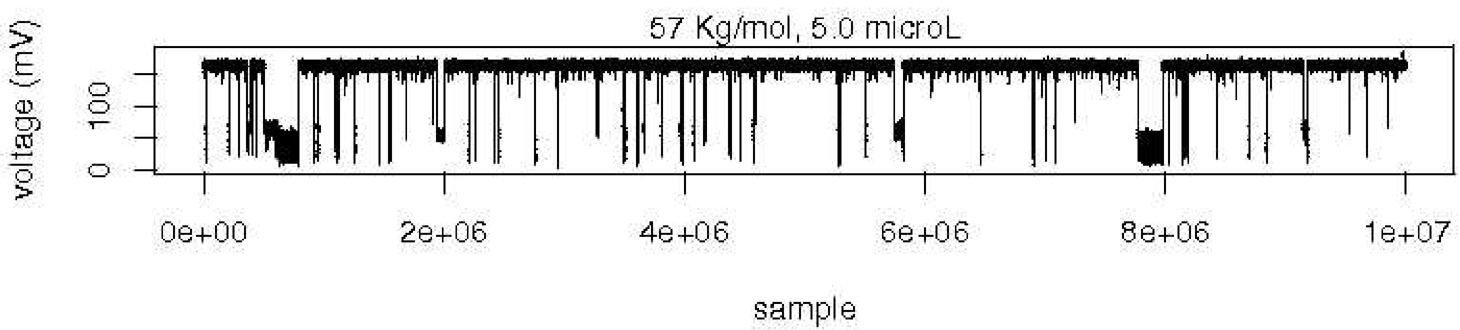


Experimental Signal of long chain polymers

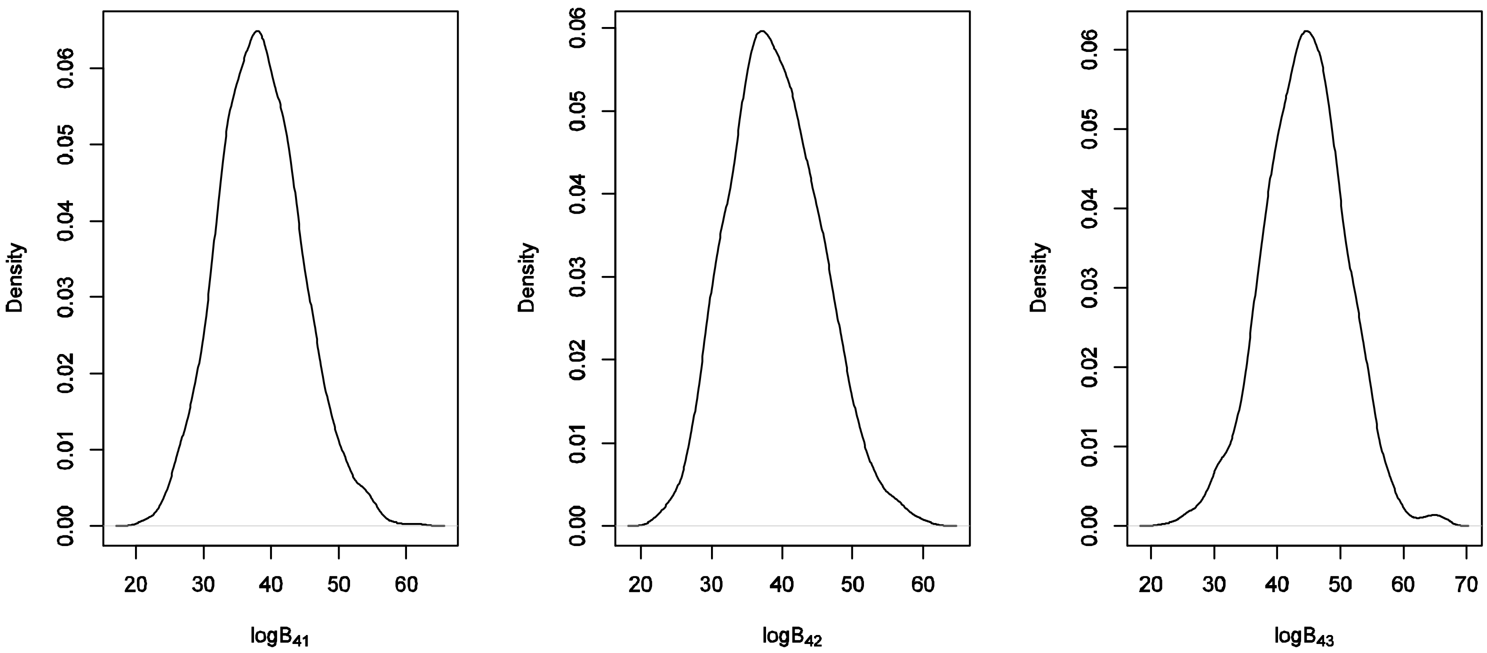


Experimental Signal as a function of the concentration





Bayes Factors Classification
Different polymers



Bayes Factors Classification
Different concentration