Comparison of DNA Analysis Using Probability State Modeling and Non-Linear Least Squares Modeling

Christopher M. Bray[1], Bo Baldetorp[2], C. Bruce Bagwell[1]

1 Verity Software House; 2 Department of Oncology, Lund University

Program Number: 208

Introduction

Analysis of single measurement DNA data has been performed by cytometrists for over 35 years using Non-Linear Least Squares (NLS) modeling (1). NLS is accurate, reproducible and accounts for population overlap. As a result, modeling of DNA data using NLS has been widely adopted as good laboratory practice (2). In cytometry, NLS has generally remained limited to a single measurement because of the complexity of modeling in correlated measurement space.

Probability State Modeling (PSM) is an analysis method that extends to any number of correlated measurements. The PSM approach has the same desirable traits as NLS, namely it is automatable, accurate, reproducible and accounts for population overlaps. If PSM performs as well as NLS, then it should be possible to develop multi-dimensional DNA models in the future.

This study compares the accuracies of PSM and NLS on synthesized data where truth is known and on real DNA histograms to investigate whether PSM is capable of extending nuclear DNA modeling to high-dimensional data.

Materials & Methods

The first part of this study examines the accuracy of both NLS and PSM against a set of generated data where "truth" is known. Three hundred data files were generated to simulate one DNA cell cycle with varying proportions, means, and measurement error. ModFit LT 3.3 was used for NLS analysis and PSM analysis was performed using GemStone 1.0.49. Results were compared to the "truth" values to determine the accuracy of each method.

The second part of this study compares PSM analysis of real DNA samples with NLS results. The data set included twenty-six fresh/frozen mammary tumor biopsies and twenty-six paraffin-embedded endometrial tissue samples. Samples were prepared and analyzed by the Oncology Department at Lund University according to their DNA analysis guidelines (3).

References

1. Dean, P. N. & Jett, J. H. (1974). Mathematical Analysis of DNA Distributions Derived from Flow Microfluorometry. *The Journal of Cell Biology*, 60, 523-527.
2. Shankey T. V., Rabinovitch P. S., Bagwell C.B., Bauer K. D., Duque R. E., Hedley D. W., … Cox C. (1993). Guidelines for Implementation of Clinical DNA. *Cytometry*, 14 (5), 427-477.
3. Baldetorp, B. Guidelines for DNA FCM Analysis. Retrieved from http://www.sfff.se/file-cabinet/DNAGUIDLINES.pdf

Synthesized DNA Data

Histograms showing ModFit LT (left) and GemStone (right) analysis of one generated sample.

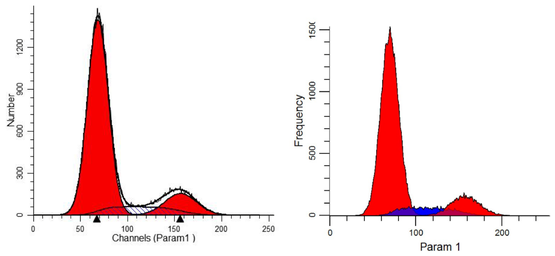


Figure 1: A comparison of NLS modeled "S-phase" to known truth for synthesized data shows nearly perfect correlation for all samples.

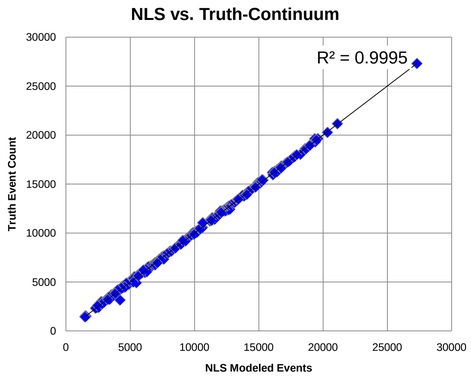
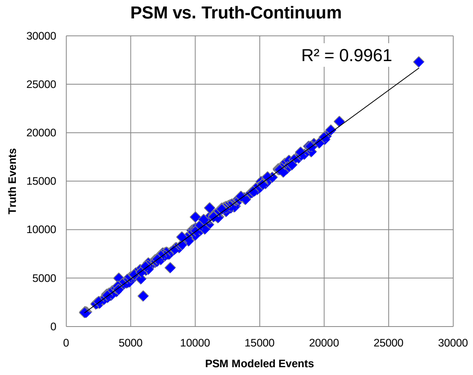


Figure 2: A comparison of PSM modeled "S-phase" to known truth for synthesized data shows nearly perfect correlation for all samples.



Real DNA Data

Histograms showing ModFit LT (left) and GemStone (right) analysis of one fresh\frozen sample.

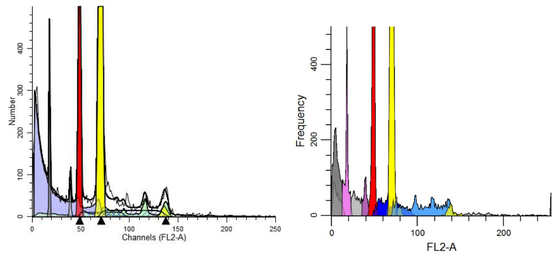


Figure 3: Correlation of total S-phase by PSM of the fresh\frozen and paraffin embedded DNA samples to the NLS analysis by Lund University.

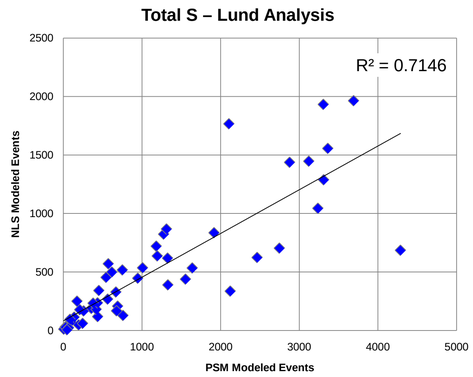
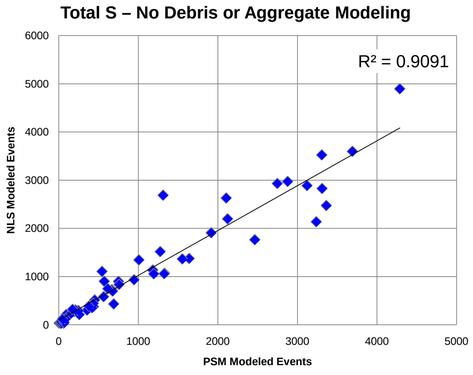


Figure 4: Correlation is greatly improved by disabling the NLS components for modeling debris and aggregates



Results

NLS results correlate well with "truth" values of the synthesized DNA data. The correlation for the percentage of S-phase was R^2=0.9995 (Figure 1). The correlations for G0/G1 and G2 were R^2=0.9999 and R^2=0.9996, respectively. Cases that had the greatest deviation from the truth values had highly-overlapping populations or populations that were partially off scale. PSM analyses of the generated data also show high correlation with "truth". The correlation of the PSM estimate for S-phase was R^2=9961 (Figure 2). PSM results had a correlation of R^2=0.9995 for G0/G1 and R^2=0.9996 for G2. PSM results with the greatest deviation from "truth" were those with highly overlapping populations. PSM did not have difficulty with the cases that were partially off scale. Both NLS and PSM performed well when analyzing simple generated data.

In the second part of the study analyzing real DNA samples, the Lund NLS analyses were presumed to be the best estimates of the "truth". Therefore, the PSM analysis was compared to the NLS results. Since the real DNA data may have more than one cell cycle, total S-phase was compared to give a better reflection of the overall correlation between the two methods. This comparison yielded a correlation of R^2=0.7146 between the two methods for the total S-phase of the samples (Figure 3). The correlation for the diploid G0/G1 was R^2=0.9254. Given that the PSM did not account for debris and aggregates, an additional comparison was made. By modifying the Lund analysis to exclude modeling of debris and aggregates, the correlation between PSM and NLS improved to R^2=0.9091 (Figure 4). For the modified NLS results, the correlation for the diploid G0/G1 was R^2=0.987. Due to differences in how the diploid G2 is modeled in tetraploid cases, the results for the diploid G2 population were not well correlated in either comparison, R^2=0.061 and R^2=0.1091 respectively.

Conclusions

- PSM and NLS have equivalent accuracy (R^2 >0.99) modeling overlapping populations of simple generated data.
- PSM models data that is partially off scale more accurately than NLS.
- PSM analysis of single measurement DNA would benefit by the addition of the ability to model debris and aggregates.
- The greatest advantage to PSM is its scalability to model any number of measurements.