ACAT 2017
21-25 August 2017
University of Washington, Seattle

The experience of comparative analysis of data in a biophysical experiment

Arturo Matamoros-Volante, Olga Bondarenko
Departamento de Genética del Desarrollo y Fisiología Molecular, IBT-UNAM, Cuernavaca, México
Elena Nurmatova, Nikita Okunev
Moscow Technological University (MIREA)
Sergey Bityukov, Vera Smirnova
SRC Institute for High Energy Physics NRC "Kurchatov Institute"

Introduction

This report presents preliminary results of the analysis of biophysical experimental data using the method of statistical comparison of histograms.

A method for statistical comparison of histograms [arXiv:1302.2651; European Physical Journal Plus, 128 (2013) 143] allows to compare experimental and / or theoretical data with the help of multidimensional test statistics, which is built on the basis of the empirical distribution of test statistics of the "significance of the difference" for corresponding bins of the compared histograms.

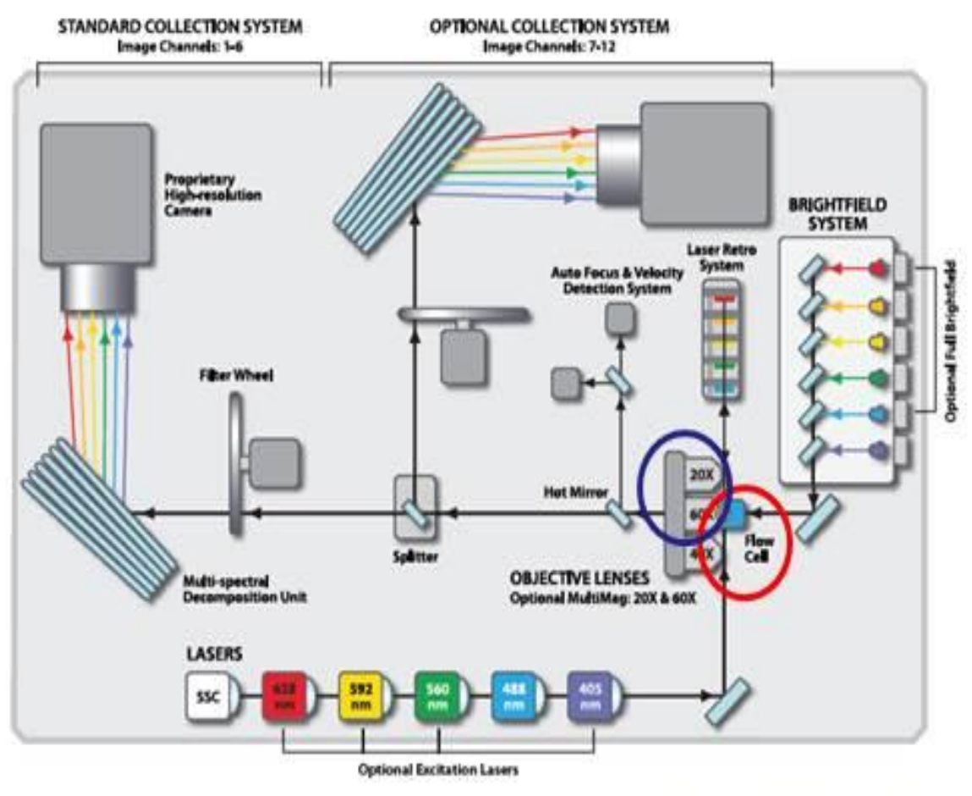
In this example, two-dimensional test statistics (**<S>, rms**) are used as a two-dimensional "distance" between histograms. Here, **<S>** is the mean value, and **rms** is the mean square of the "significances of the difference" distribution.

Measurement model I

- A fluorescent label **(Disc$_{(3)}$5)** was used to record the membrane potential of cells.
- The change in fluorescence was measured with an **Amnis®** instrument using a 658 nm laser.
- This device allows simultaneous registration of cell fluorescence in 2 modes:

1. Flow Fluorimetry:
The cells pass one after the other through the laser beam, the fluorescence and the scattered laser radiation of each cell are registered by the detector.

2. Fluorescence microscopy:
The image of the passing cells is recorded by fluorescence microscopy using a high resolution video camera.


STANDARD COLLECTION SYSTEM
Image Channels: 1-6
OPTIONAL COLLECTION SYSTEM
Image Channels: 7-12
Proprietary High-resolution Camera
Filter Wheel
Multi-spectral Decomposition Unit
Splitter
Auto Focus & Velocity Detection System
Laser Retro System
BRIGHTFIELD SYSTEM
Optional Full Brightfield
Hot Mirror
20X
60X
40X
Flow Cell
OBJECTIVE LENSES
Optional MultiMag: 20X & 60X
LASERS
SSC
642 nm
592 nm
560 nm
488 nm
405 nm
Optional Excitation Lasers

The structure of the device

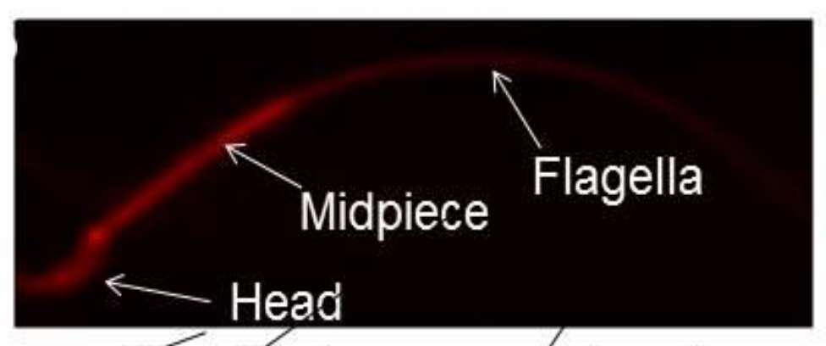
Results of measurements are displayed on the screen in the form of histograms and cell images

Measurement model II

In this experiment, a signal is registered from 5000 cells.

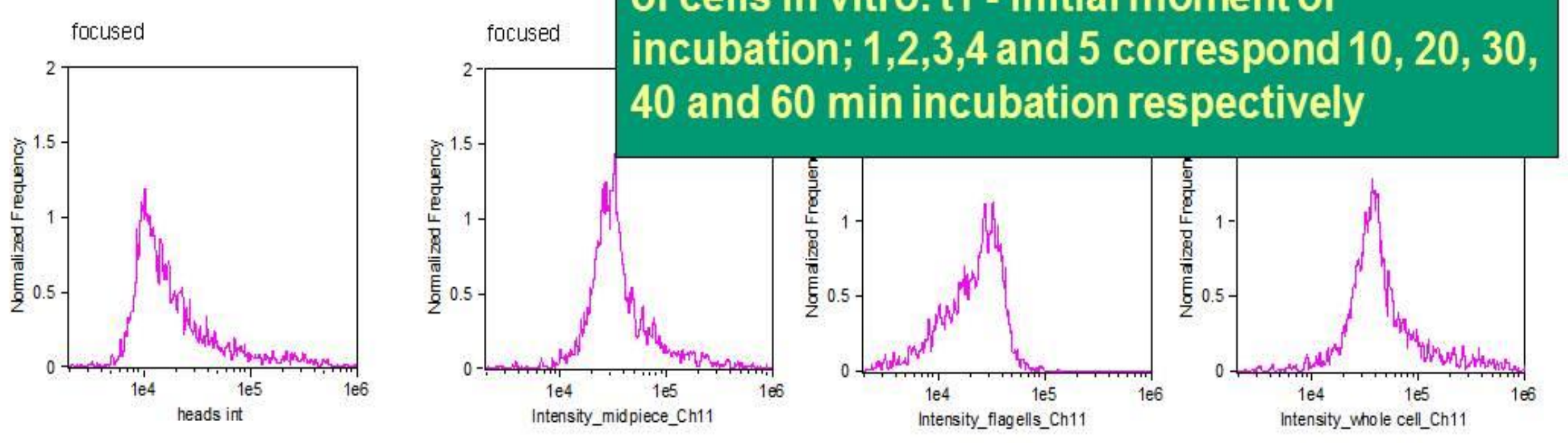
With the help of filters from 5000 cells, only those that are in focus at the time of registration are selected. The cells that are not in focus are skipped. Thus, a finite number of cells for analysis is about 2-2.5 thousand.

Cell in focus



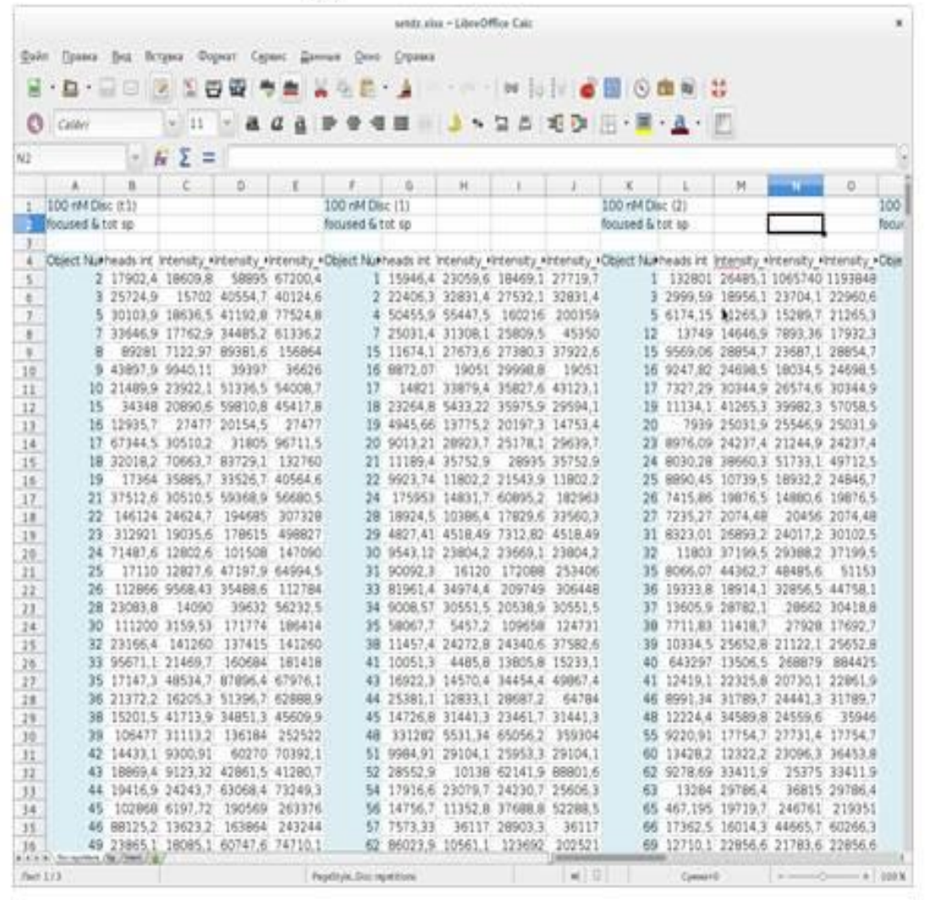
The use of some masks allows you to record the intensity of fluorescence in various areas of the cell (heads, Intensity_flagella, Intensity_midpiece, Intensity_whole cell).
The histograms of fluorescence intensity of cells or selected regions of cells is displayed on the screen of the device.

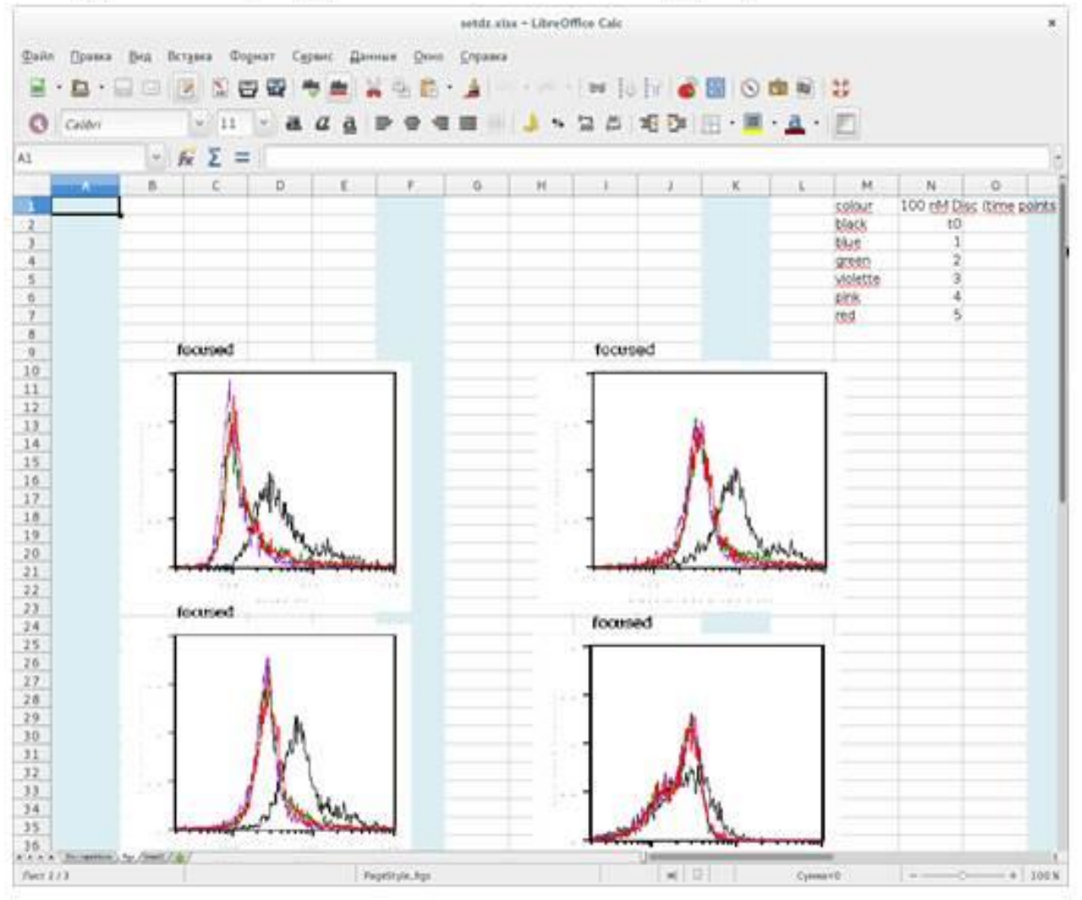
The goal of experiment: to study the change in membrane potential during short-term storage of cells in vitro. t1 - initial moment of incubation; 1,2,3,4 and 5 correspond 10, 20, 30, 40 and 60 min incubation respectively



Measurement model III

The data comes from the device in the form of an Excel file in the form of a table, part of which is shown below (table on the left), and in the form of 4 sets of histograms for each of the cell regions (figures on the right).





1, 2, 3, 4 and 5 correspond 10, 20, 30, 40 and 60 min incubation respectively, t1 - initial moment of incubation.

The significance of difference I

Let take the two histograms Disc(t1) and Disc(1) for the head area with the number of bins n=500 (the figure in the upper left on slide # 8). Suppose that the contents of the bins are

Disc(t1) : $\hat{n}_{11}\pm\hat{\sigma}_{11}, \hat{n}_{21}\pm\hat{\sigma}_{21}, \ldots, \hat{n}_{n1}\pm\hat{\sigma}_{n1}$;
Disc(1) : $\hat{n}_{12}\pm\hat{\sigma}_{12}, \hat{n}_{22}\pm\hat{\sigma}_{22}, \ldots, \hat{n}_{n2}\pm\hat{\sigma}_{21}$, где $\hat{\sigma}_{ij}=\sqrt{\hat{n}_{ij}}$ (common assumption in many tasks).

Then the significance of the difference for bin # *i, i = 1,n* is determined by the formula $\widehat{S}_i=\frac{\hat{n}_{i1}-K\hat{n}_{i2}}{\sqrt{\hat{\sigma}_{i1}{}^2+K^2\hat{\sigma}_{i2}{}^2}}$, where *K* – normalization factor (usually this is the ratio of the volumes of samples studied). If the samples are independent and taken from the same population, then each value $\widehat{S}_i$ is the realization of a random variable close to the standard normal distribution *N(0,1)*. This means that the distribution of **n** values of S_i is also close to *N(0,1)*.

The significance of difference II

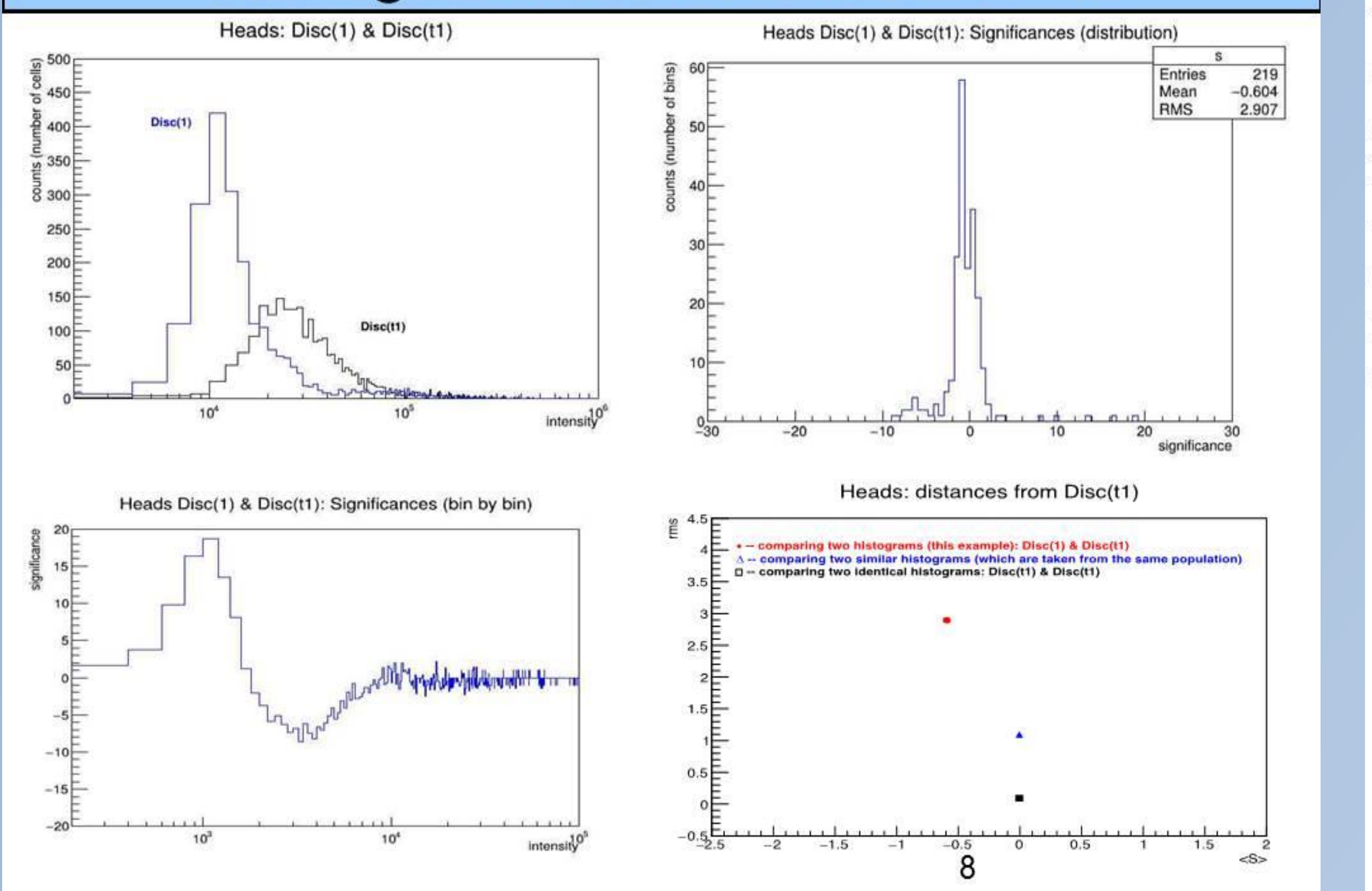
As a result, on the basis of empirical distributions of both the bitwise and the total distribution of values, it is possible to analyze the compatibility or distinguishability of the samples studied.
In the example under consideration, two statistical moments of the distribution are used: the mean value of significance $\bar{S}$ and the mean square *rms*, that is, the two-dimensional test statistics

$$SRMS = (\bar{S}, rms).$$

If *SRMS* = *(0,0)*, then the histograms are identical,
if *SRMS* ≈ *(0,1)*, then the original samples are taken from the same population.

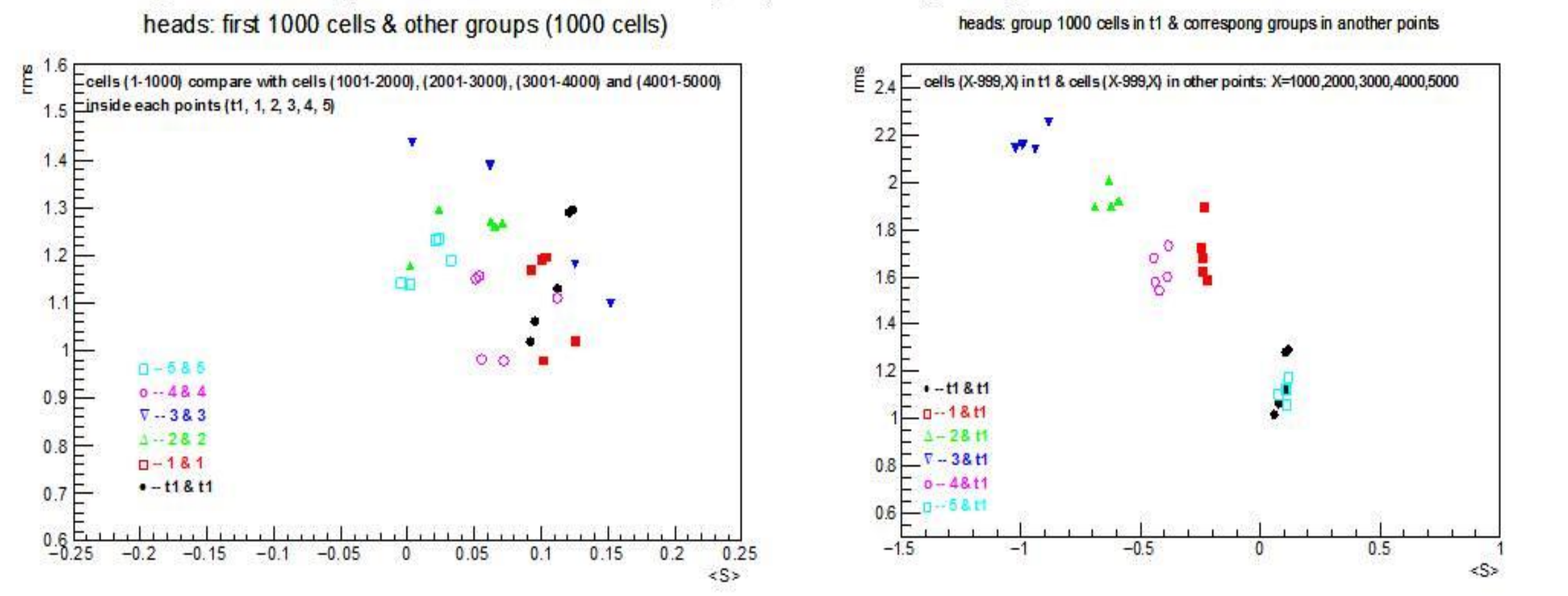
In this example, the total distribution is used (the upper figure on the right on the next slide). The bitwise distribution is shown in the figure on the lower left. The corresponding *SRMS* values are given in the figure at the bottom right.

The significance of difference III



The significance of difference IV

To assess the ability of the instrument to distinguish between sets of data, the data of each measurement was divided into 5 partitions by cell number. A histogram was constructed for each partition. The histograms were compared. In the left figure, the histogram comparison results for cells from the first to 1000 with the histograms of the other parts of each measurement. In the right figure, a comparison of the histograms of each part of the cells in the t1 measurement with the corresponding histogram of other measurements. It can be seen that the device distinguishes between test data. There is seen insignificant systematics uncertainty (left histogram).



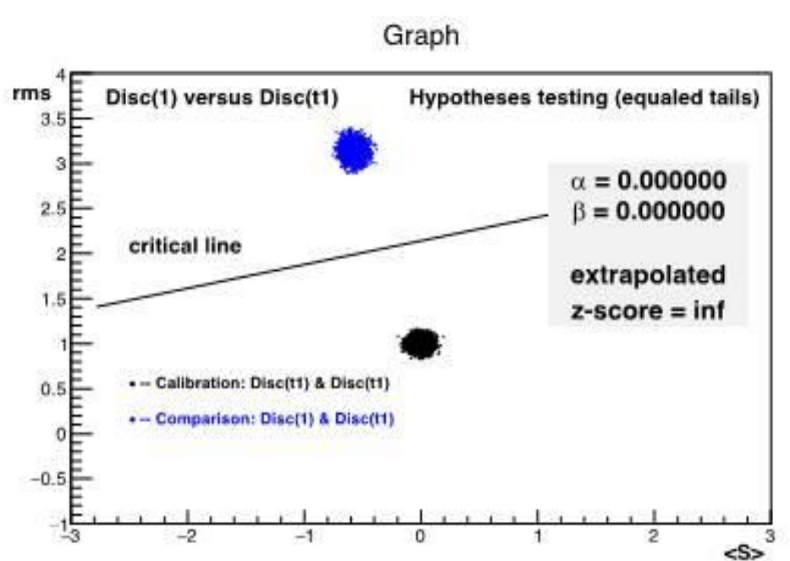
Hypothesis testing I

If the purpose of comparing histograms is to determine whether samples from which histograms are obtained are of the same population or different, then the problem reduces to testing hypotheses: **H0** - main (samples are taken from one population) and **H1** - alternative (samples are taken from different populations). But in order to test the hypotheses, we need to estimate the uncertainty in the accuracy of the measurement, which depends on the number of bins, the normalization coefficient ***K***, and a number of other parameters, that is, we need to construct a confidence distribution for the measured value ($\bar{S}$,*rms*) and for a calibration measurement characterizing the distribution of test statistics *SRMS* if both samples are taken from the same population. hen we need to determine the critical region, which allows us to estimate the error of the Type I ***α*** the error of the Type II ***β***.

The Type I error ***α*** is the probability of making a mistake when choosing hypothesis **H1**, if **H0** is true, the Type II error ***β*** is the probability of making a mistake choosing **H0** if **H1** is true.

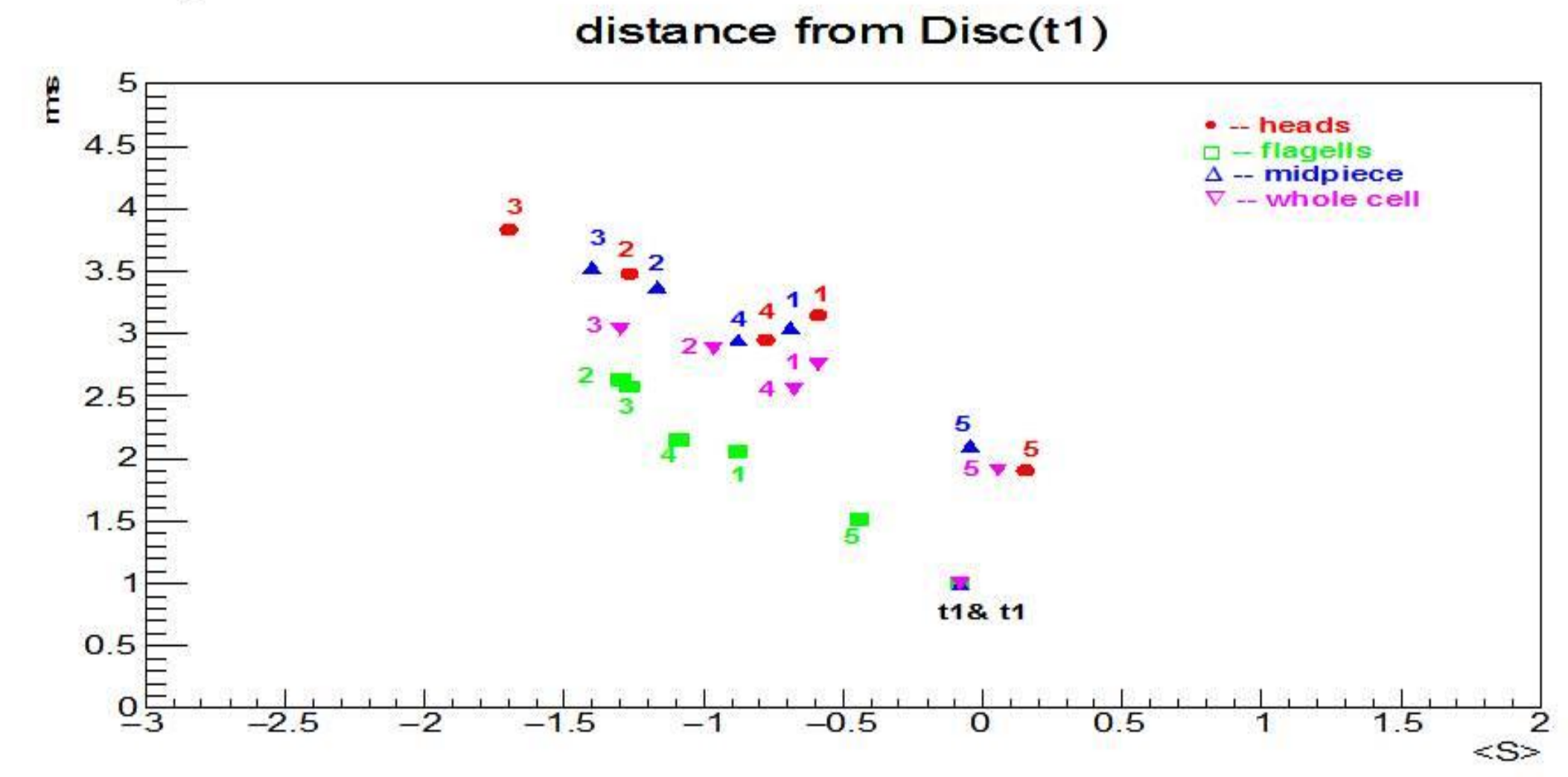
Hypothesis testing II

By choosing the significance level ***α***, one can estimate the power of the criterion (the choice of the critical region) ***1 – β***. Usually the significance level is chosen equal to 10% or 5%. It is more convenient to introduce a probability distance, either in the form of relative uncertainty
(α + β)/(2 - (α + β)), where ***α + β ≤ 1***, or as the mean error ***(α + β)/2***. In our example, the distribution of confidence in the compared quantities is constructed by the method of a repeated histogram, that is, similar histograms similar to the original are modeled according to the values and errors in the bins of the original histograms. The figure shows the confidence distributions for the test (blue spot) and calibration (black spot) measurements. In this example, there is no need to estimate the probabilistic distance, since the histograms are 100% distinguishable.



The result of the test analysis

A comparison was made of the histograms obtained at the initial time t1 and the data obtained at time points 1, 2, 3, 4 and 5, for different regions of the cell. The figure shows the two-dimensional distances between the corresponding histograms. There is a correlation between the measurement number and the intensity of fluorescence for different regions of the cell.



Conclusions

- In this report is considered the possibility of using the method of statistical comparison of histograms for analyzing the dynamics of biophysical processes in cellular organisms.
- It is shown that this approach allows to monitor and visualize changes in biophysical processes at the cellular level during mass data processing.
- When using the bitwise distribution of the significance of the difference, a deeper analysis of the initial data is possible.

Instituto de Biotecnología
UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

МИРЭА

ИФВЭ