UNIVERSITÄT LEIPZIG

n-dimensional segmentation of heterogeneous data

Halima Saad'Allah SAKER

Motivations

- Comparing changes in genomic organization
- Identification of functional units on the genomic DNA that behave coherently in multiple conditions and tissues
- **Multi-dimensional** segmentation of multivariate genomic/epigenomic/proteomic data from multiple time points/tissue/cell types, development stage …

Motivations

- Segmentation of genomes into limited number of element types using a large collection of heterogeneous annotated data tracks as input
- Identify the “thin segments” on which the values of interval in all signals are quite constant

Introduction

- The main goal is to **design**, **implement**, and **test** novel segmentation algorithms that work on one- and multidimensional
- Algorithm can accommodate data of different types and resolution
- We don't want to just implement a general purpose segmentation algorithm but first and foremost to design one

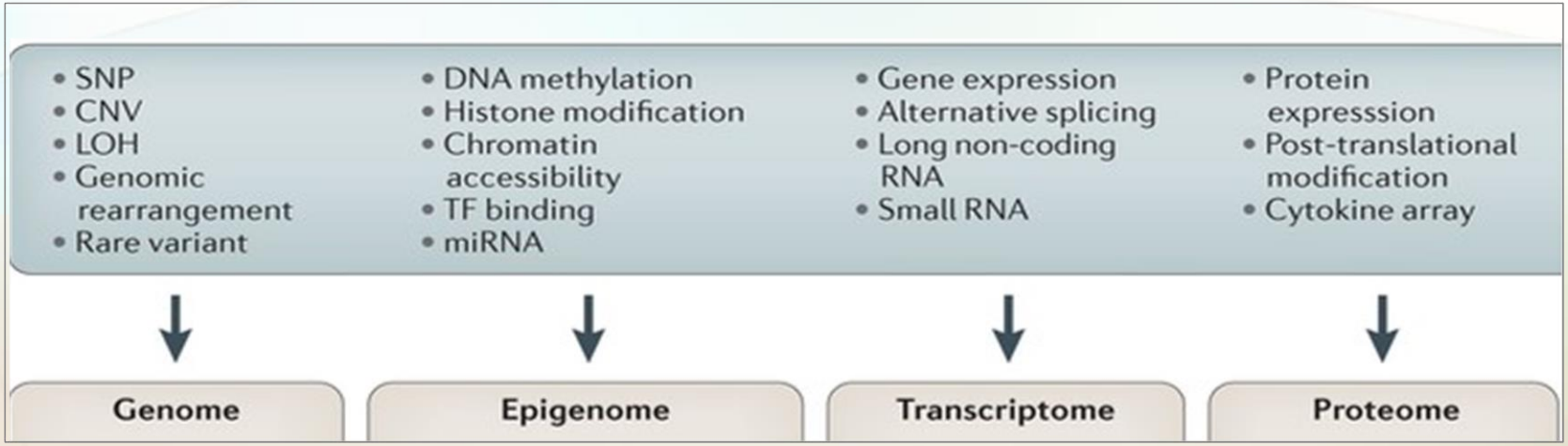
Introduction

- The idea is that a segment, a gene or processed part of a gene such as an intron, have everywhere the same expression levels over time
- Should be applied to integration of heterogeneous data to improve the map of functionally coherent segments of a genome, in particular of course the human one

GENOME SEGMENTATION

- The goal of segmentation in Bioinformatics is to decompose the genomic sequence, into a small number of **homogeneous** non-overlapping pieces, **segments**.
- Each **segment** has a certain degree of internal *similarity*.

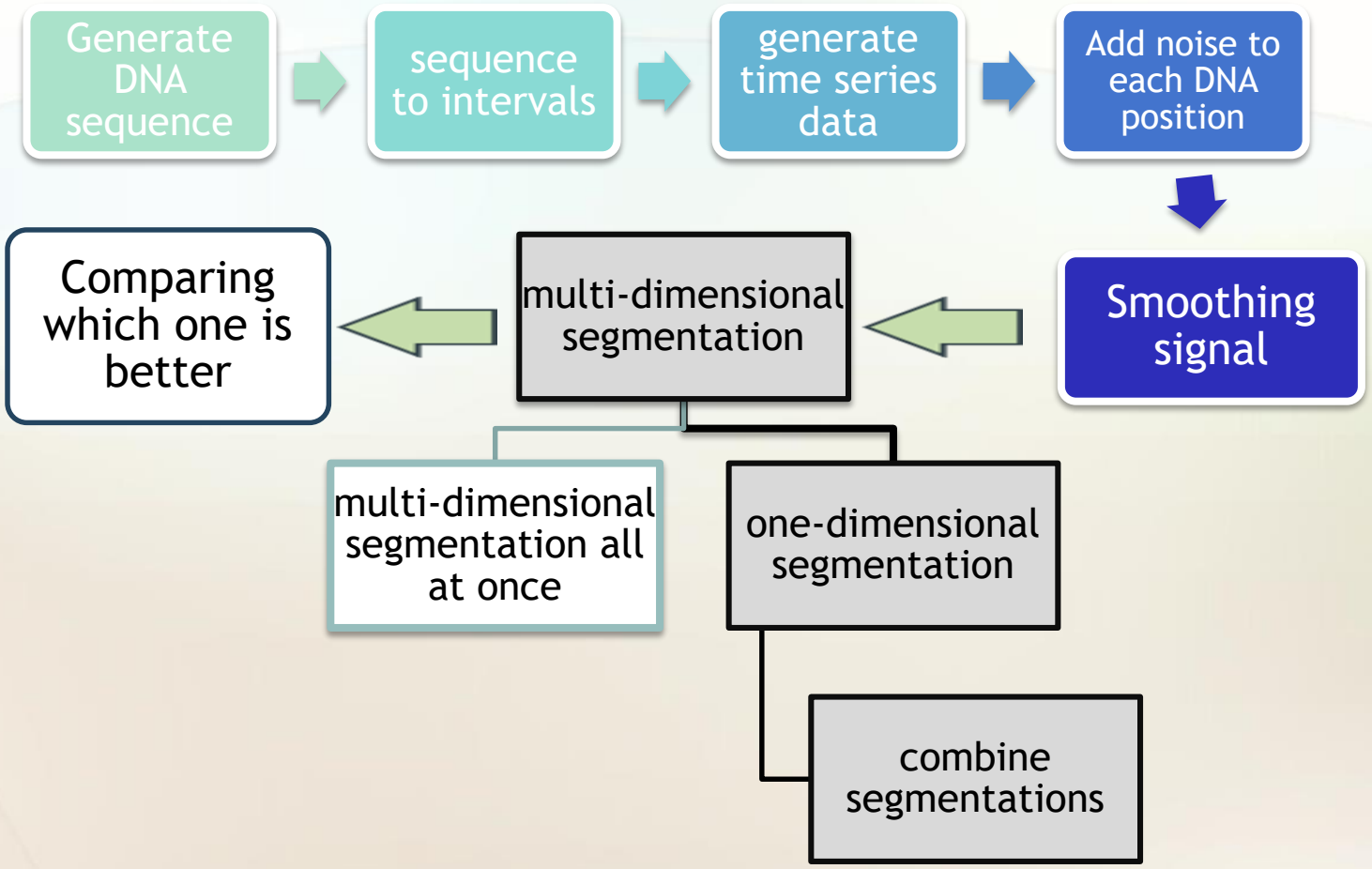
Data types


• SNP
• CNV
• LOH
• Genomic rearrangement
• Rare variant
• DNA methylation
• Histone modification
• Chromatin accessibility
• TF binding
• miRNA
• Gene expression
• Alternative splicing
• Long non-coding RNA
• Small RNA
• Protein expresssion
• Post-translational modification
• Cytokine array
Genome
Epigenome
Transcriptome
Proteome

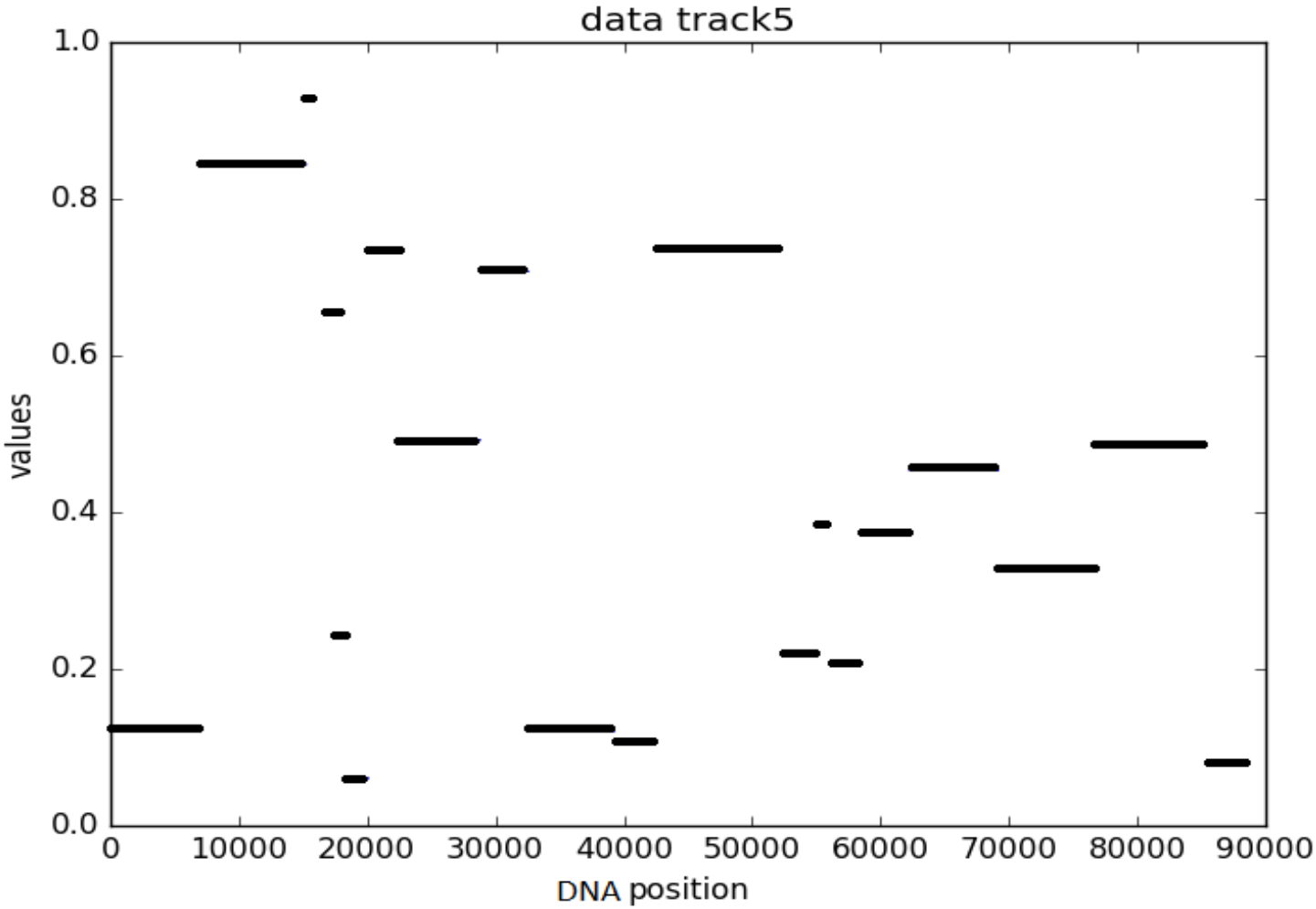
Segmentation algorithms

- Preexisted segmentation algorithms:
- HMM
- HSMM
- SegWay based on DBN
- Top down algorithm
- Bottom down algorithm
- Sliding window
- …

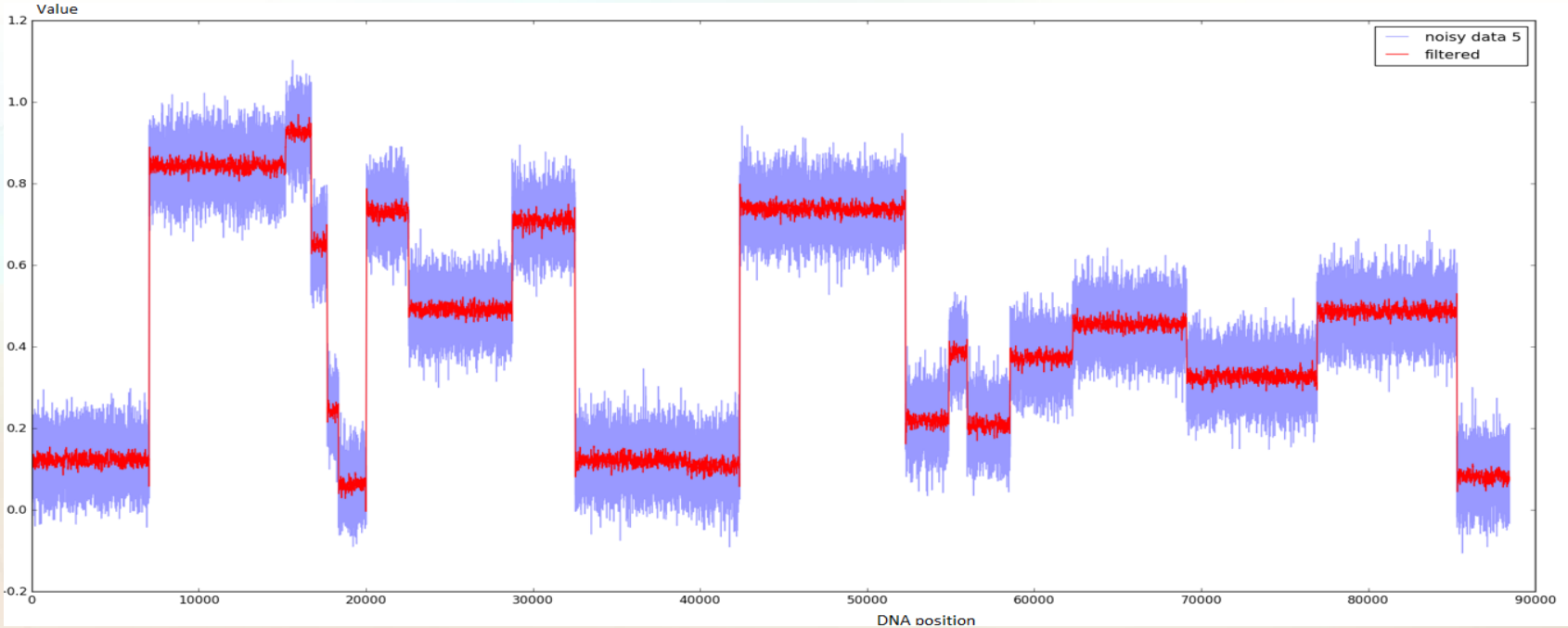
System Architecture



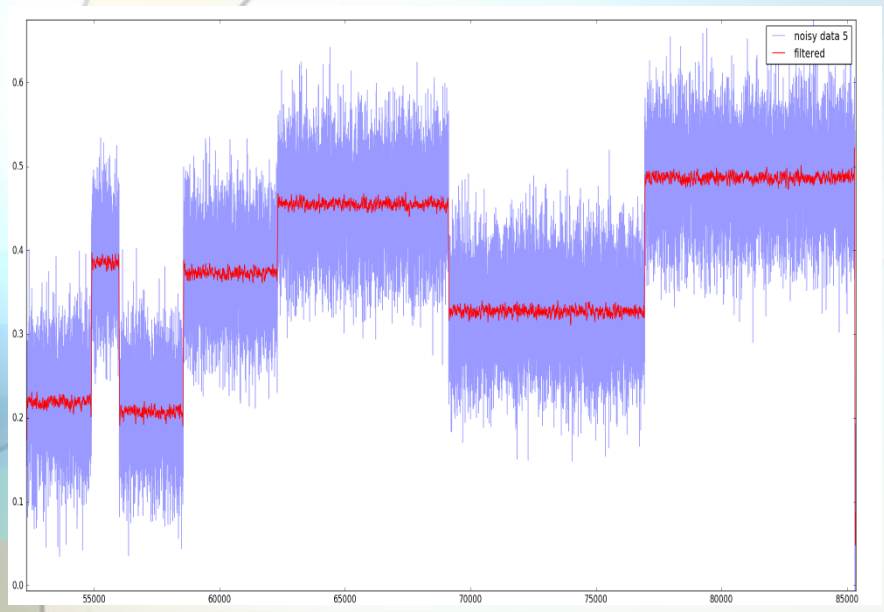
Generate artificial data: intervals value

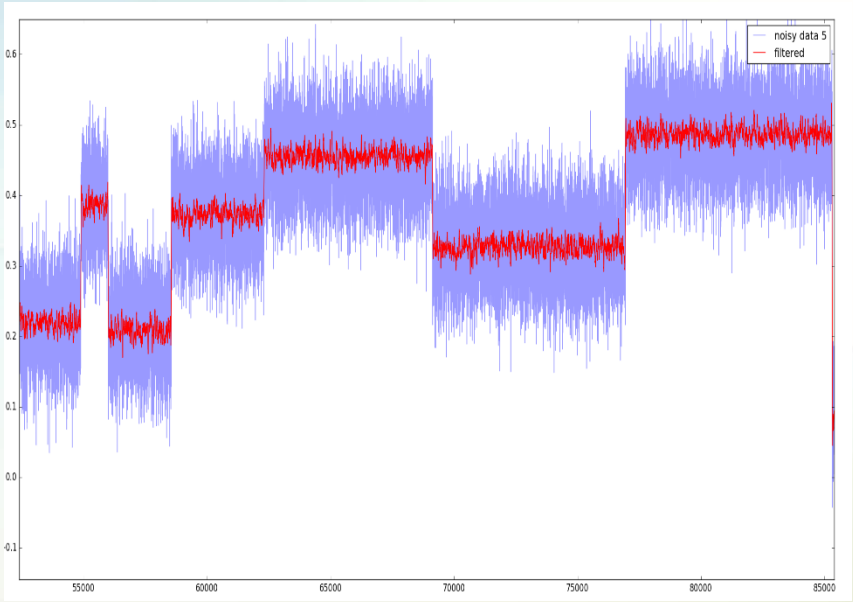


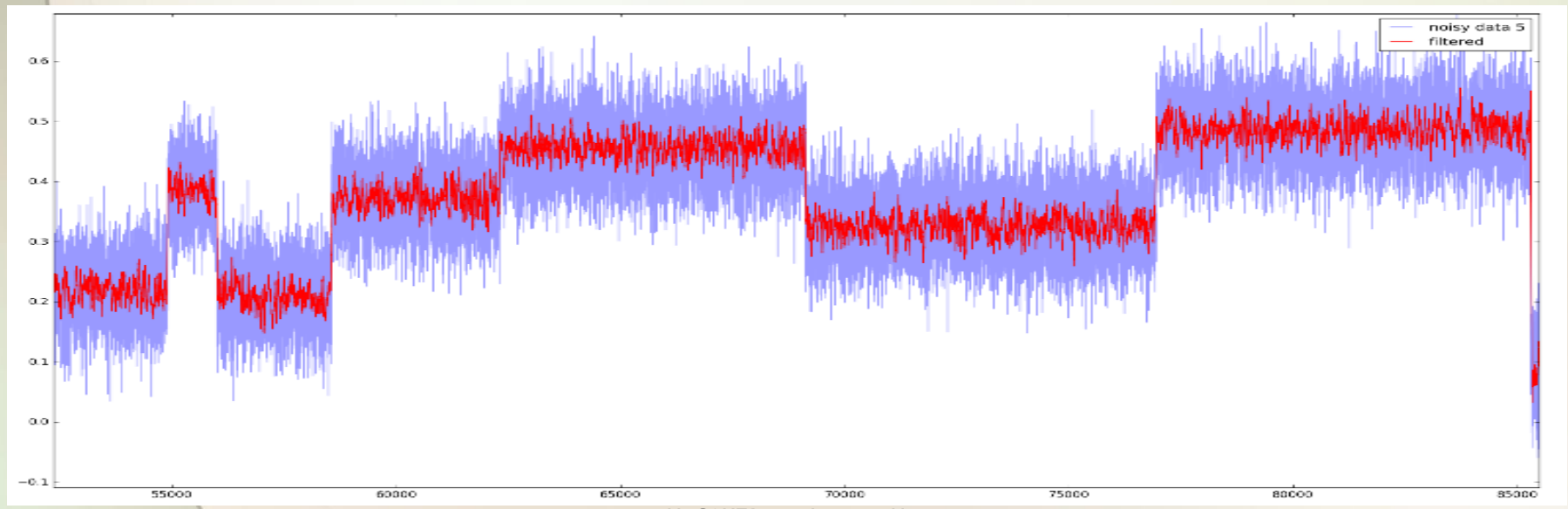
Generate artificial data: Add noise & smoothing



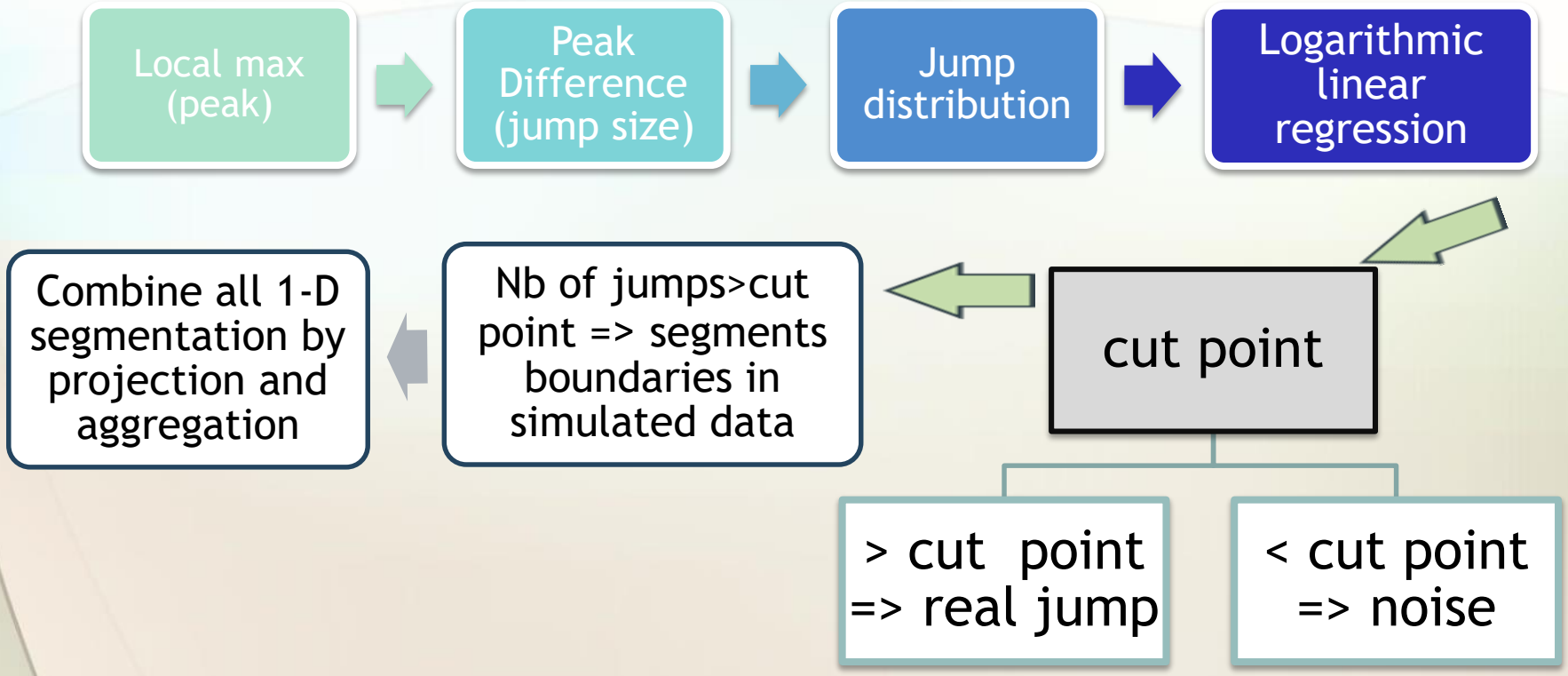
Different noise level



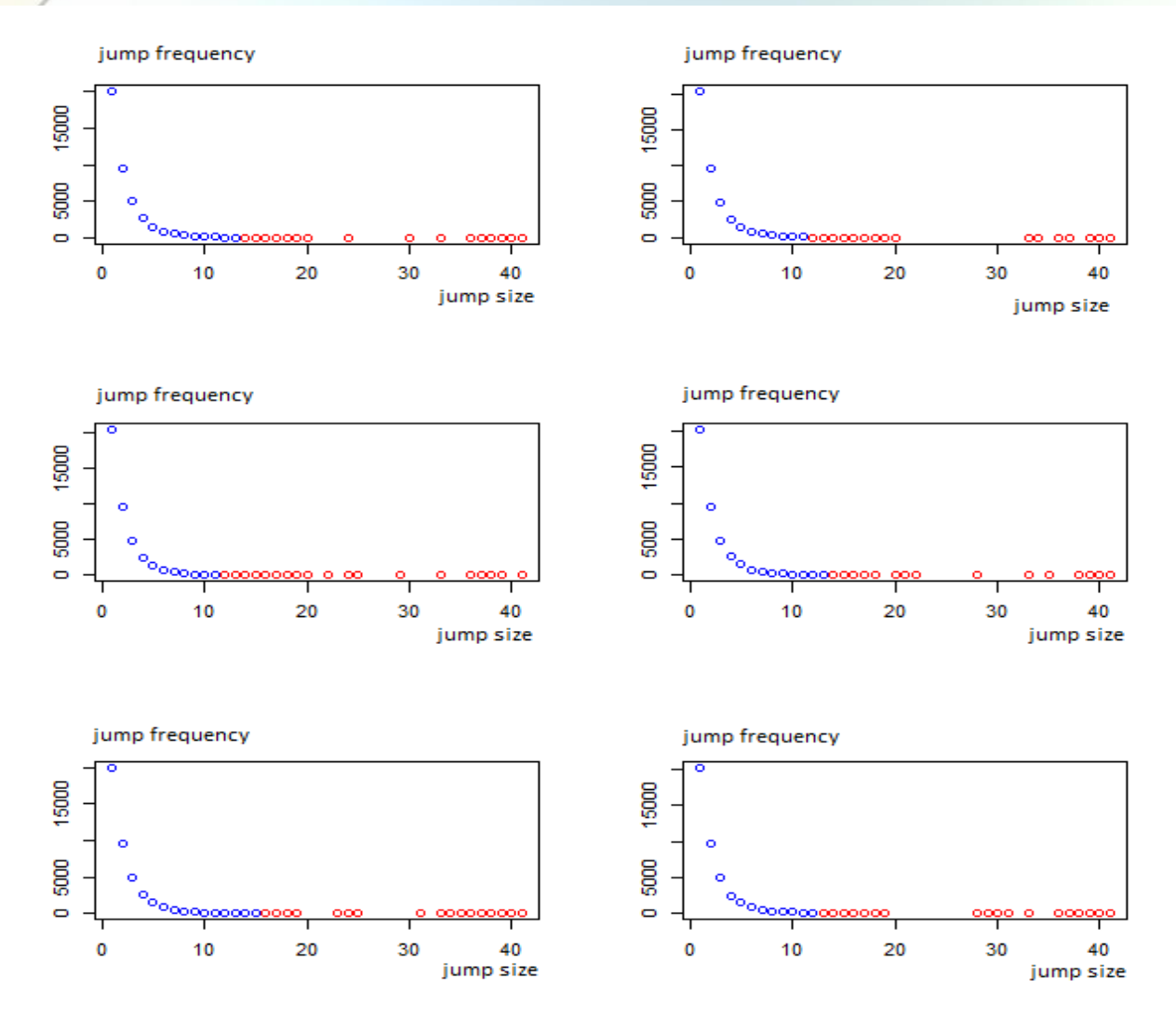




Algorithm process



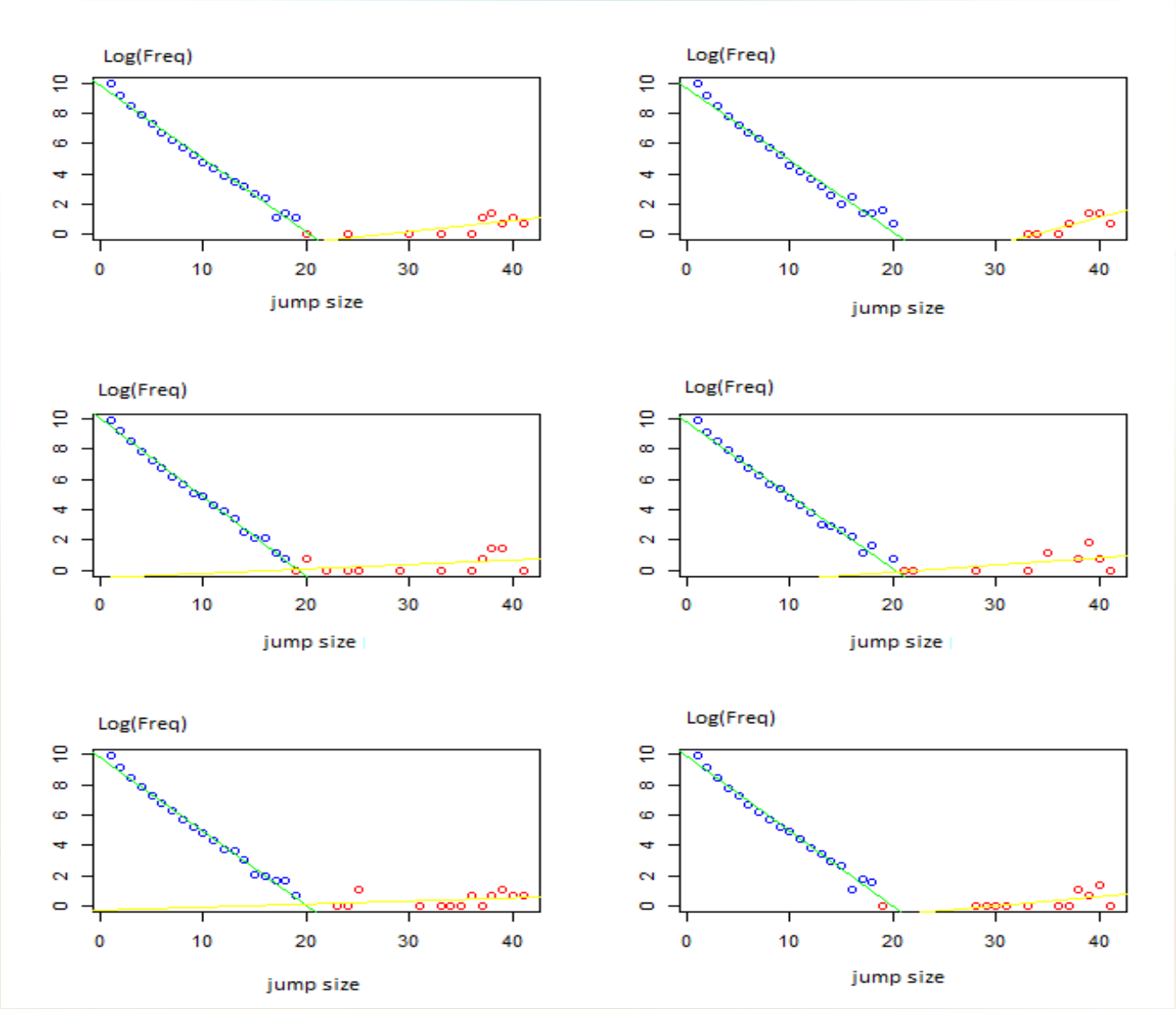
Jump distribution



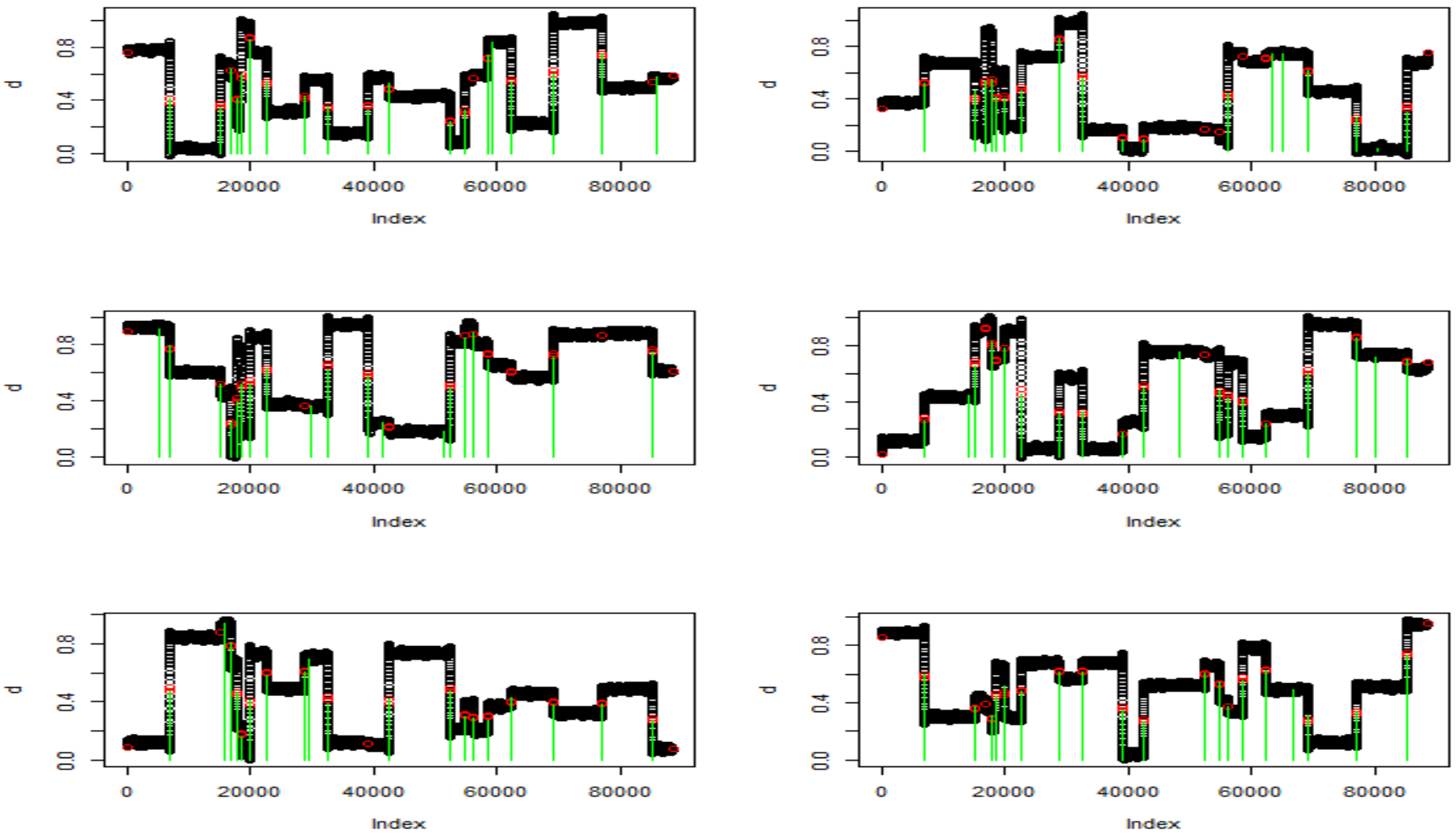
Real jump

Noise

Logarithmic linear regression: Cut point



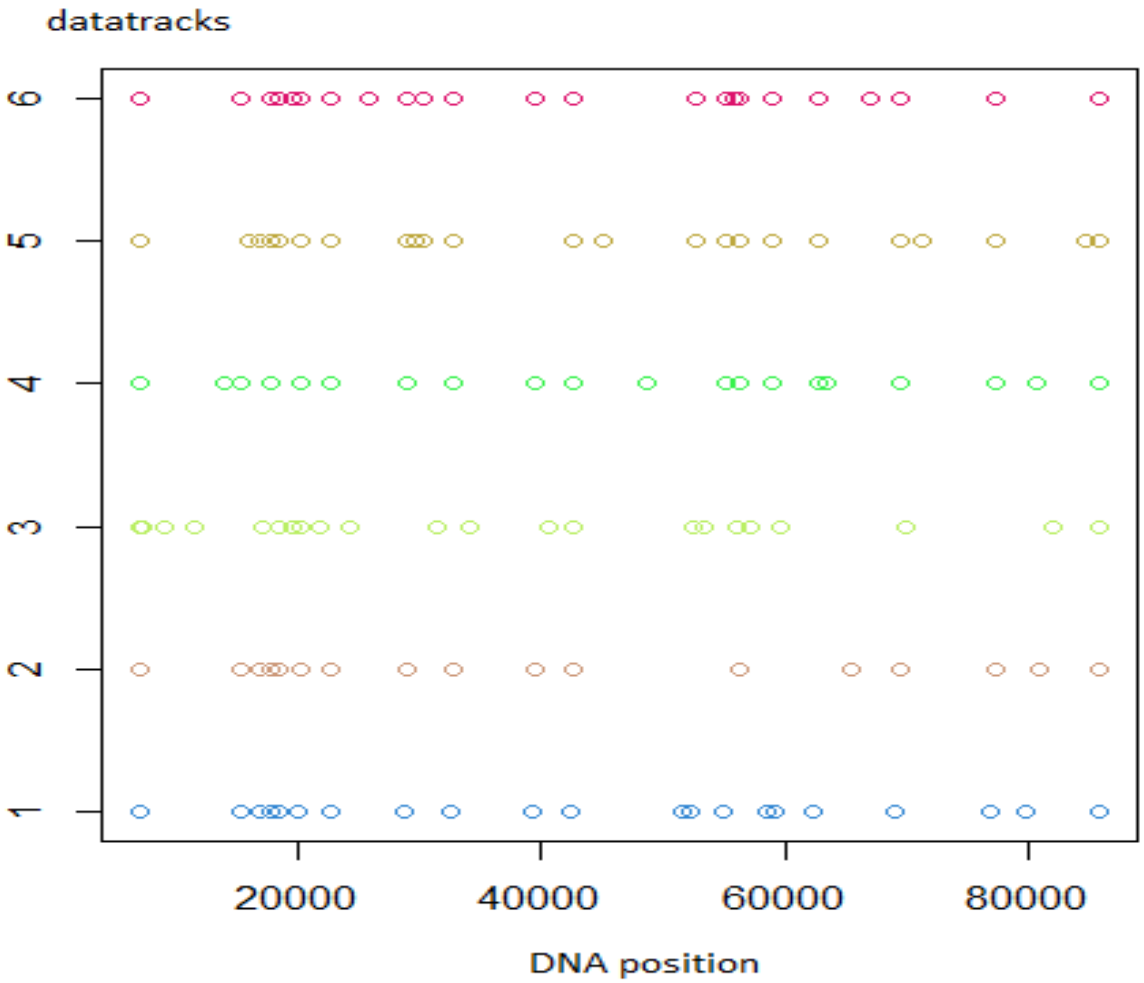
Results



Real Segment boundaries

boundaries found

Segment boundaries in each data_track



What's next!

- Combine all one dimensional segmentation by projection and aggregation
- sub-problem in segmentation is to decide whether two adjacent intervals should form two distinct segments or whether they should be combined to a single one
- merging thin segment to neighbors

Thank you!!