UNIVERSITÄT LEIPZIG

MULTIDIMENSIONAL SEGMENTATION OF SACCHAROMYCES CEREVISIAE DATA

Halima Saad'Allah SAKER

Supervisor:

Prof. Peter F. Stadler

Prof. Ahmad Shahin

OUTLINE

- Motivation
- Introduction
- Methodology
- Result
- Conclusion

MOTIVATION

- 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”.
- The main goal is to **design**, **implement**, and **test** novel segmentation algorithms that work on one- and multi data dimension.
- Segmentation of genomes into limited number of element types using a large collection of heterogeneous annotated data tracks as input.
- Identification of functional units on the genomic DNA that behave coherently in multiple conditions and tissues.

INTRODUCTION

- Segmentation algorithms are widely used for extracting regions that behave homogeneously from sequences or time series.
- With the rapid availability of high throughout data sets, the segmentation problem appears more and more in the setting of multi-dimensional data.
- The segmentation problem, which addresses the task of subdividing an ordered sequence of data into homogeneous, approximately constant intervals
- We suggest a new segmentation method based on decomposition thresholding, and local optimum differentiation
- Detects significant breakpoints in the data to identify segment boundaries

METHODOLOGY

- Each data track is segmented independently.
- The interval are then have to be reconciled in a second, independent step (combining 1-D segmentation).
- We consider the distribution of occurrence of boundaries in each 1-D segmentation
- Combining 1-D segmentation of boundaries taking into consideration the maximum segment length in each data type.

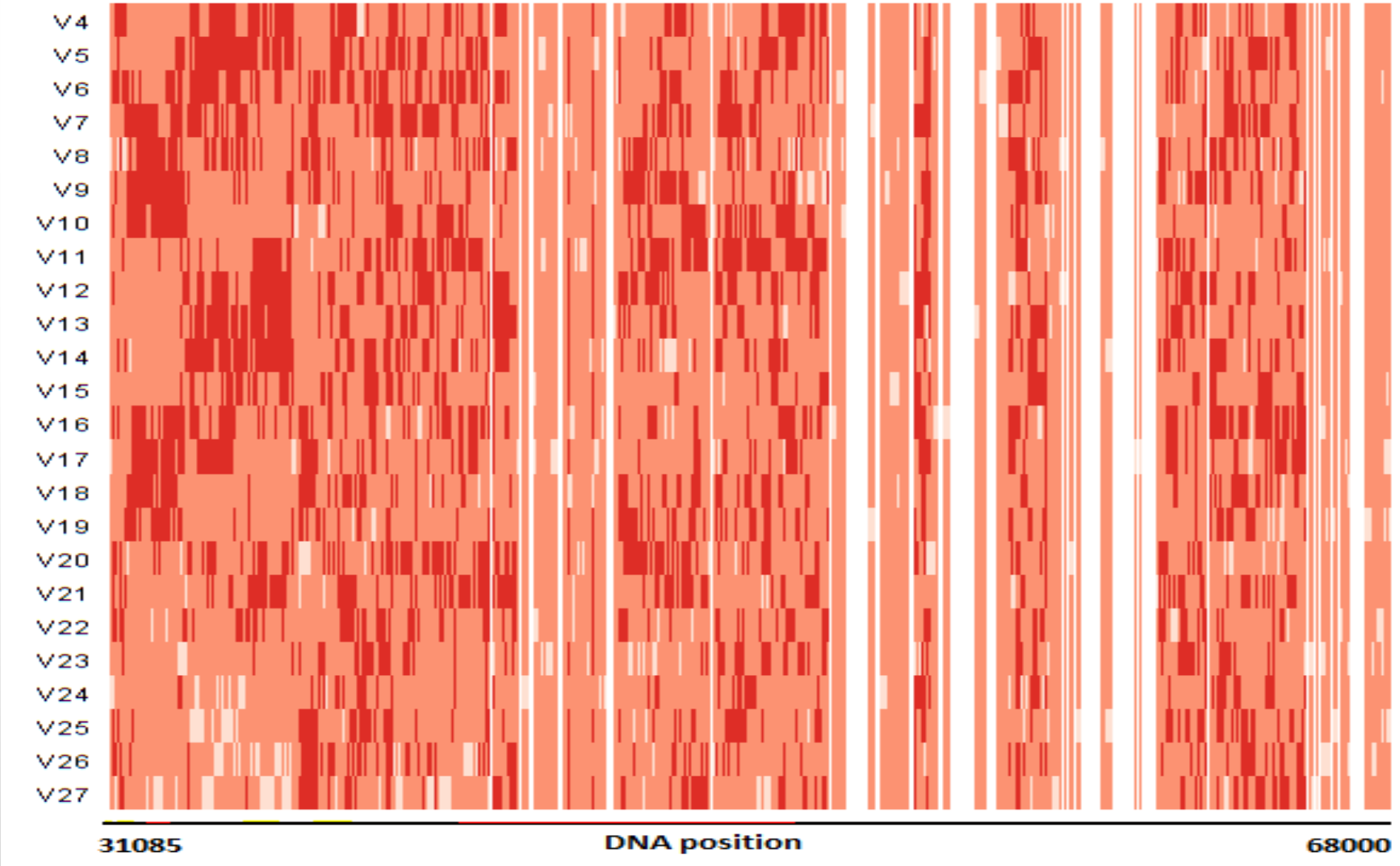
CASE STUDY

- Dataset of 24 time series transcriptomic data of Saccharomyces cerevisiae
- Data of transcribed domain from chromosome I (chrI:31100..68000)

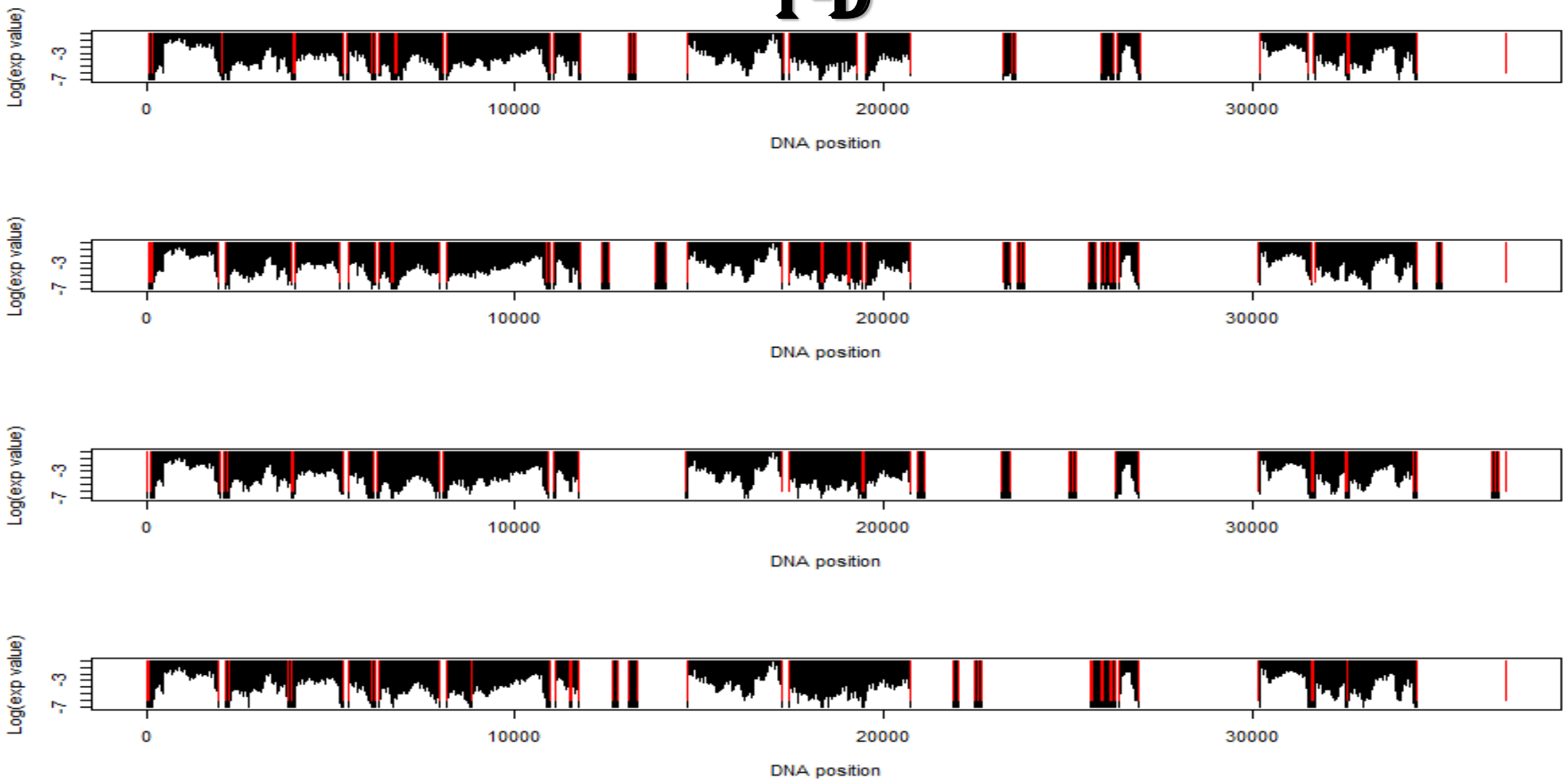
RESULTS

- The distribution of jump sizes can be used to determine significant interval boundaries in a signal independently for each data dimension.
- Segment boundaries for each datatrack are identified using the simple one-dimensional segmentation algorithm.
- These segment boundaries can then be combined to a multi-dimensional segmentation.
- The accuracy of the segmentation increases with the number of data dimensions.

HEATMAP

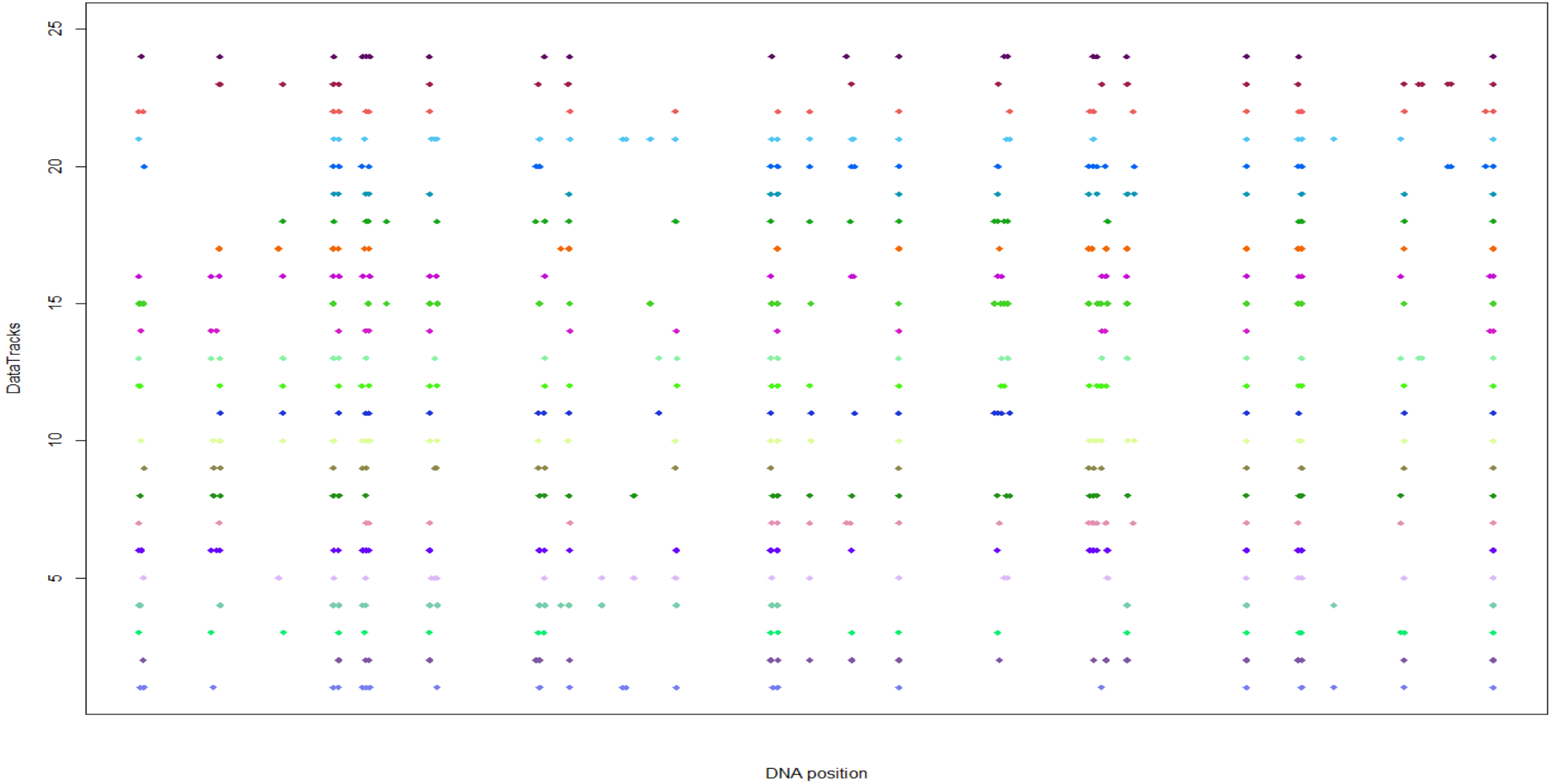

V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
V20
V21
V22
V23
V24
V25
V26
V27
31085
DNA position
68000

1-D

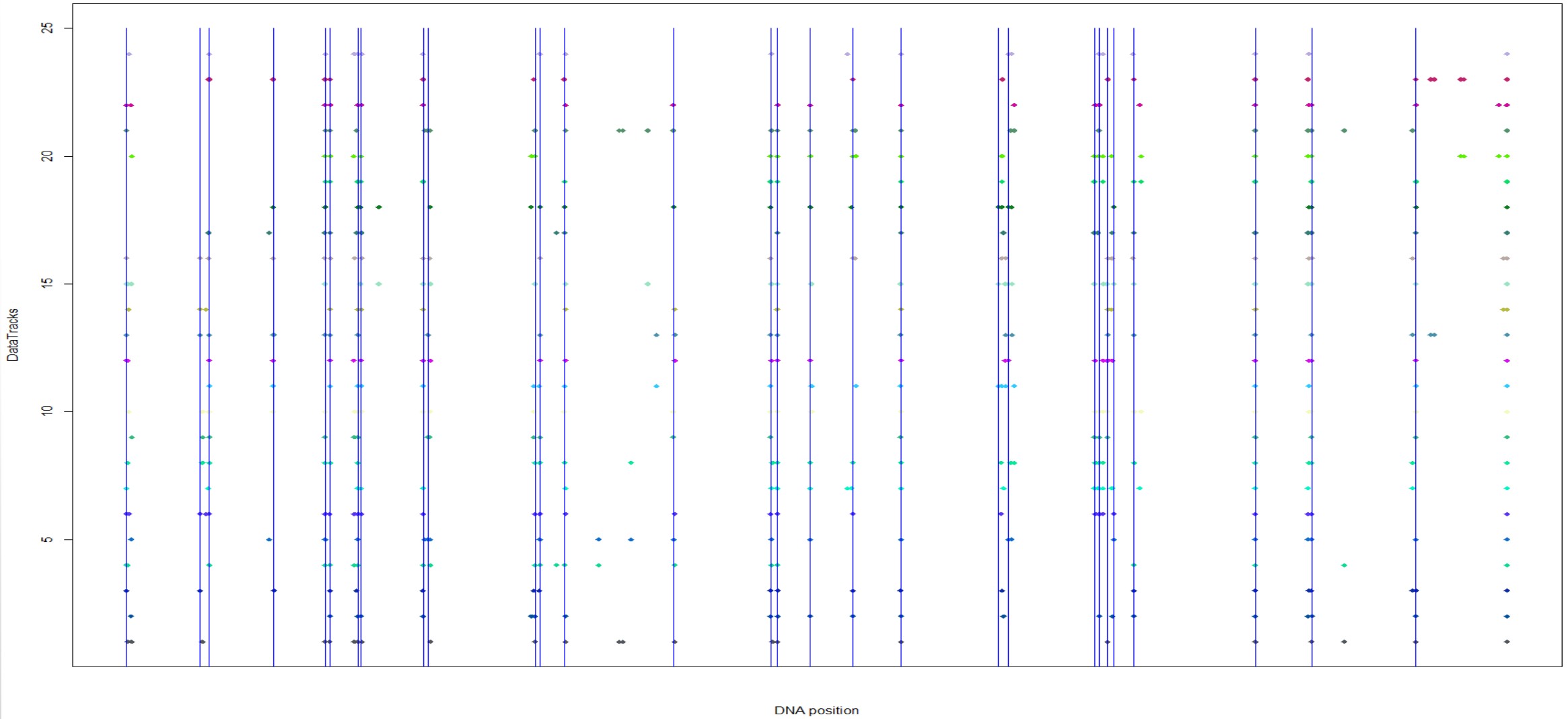


boundaries

ONE-DIMENSIONAL SEGMENTATION



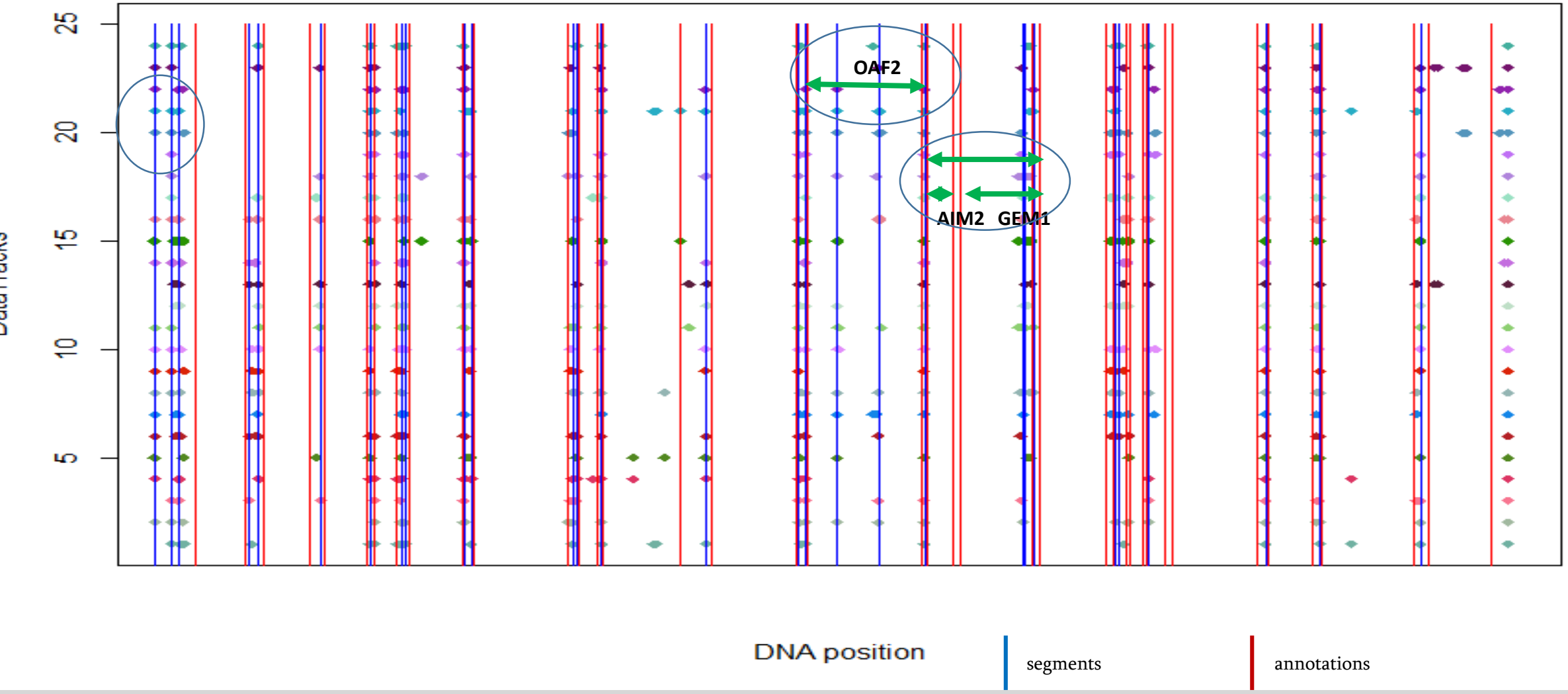
COMBINING 1-D SEGMENTATION



OBSERVATION!

- After comparing result with genome annotation, we have 3 interesting cases in our results:
  - New non annotated segment
  - one annotated gene segmented into more than segment
  - 2 or more genes have no boundaries and merged into one segment

SEGMENTATION VS. ANNOTATION



CONCLUSION

- We have presented a conceptually simple scheme for segmenting multi-dimensional transcriptomic data.
- The target data in this work is multivariate genetic/epigenetic data. The reason is that those datasets can change under the effect of several conditions such as, chemical, genetic and epigenetic modification.
- Algorithm can accommodate data of different types and resolution
- The aggregation of the 1-D boundaries leads to the desired multi-dimensional segmentation.

TO BE DONE

- Dynamic programming of algorithm to be adapted to almost all data types
- Apply algorithm on dataset of new species!
- Find new annotation in Yeast genome
- Know the biological meaning of each case!

TO BE COLLECTED

- New time series Dataset (genomic, epigenomic, transcriptomic, proteomic data) for yeast Saccharomyces cerevisiae
- Time series Dataset for other species!

ACKNOWLEDGEMENTS

- Prof. Peter Stadler
- Prof. Ahmad Shahin
- And YOU

UNIVERSITÄT LEIPZIG

MERCI POUR VOTRE ATTENTION!

Your questions, remarks and suggestions are welcome!