vl·e virtual laboratory for e·science

SigWin-detector

M.A. Inda[1], M. Roos[1], A. Belloum[2], D. Vasunin[2] and T.M. Breit[1]
Integrative Bioinformatics Unit and Institute of Informatics, Faculty of Science, University of Amsterdam, The Netherlands

Introduction

Virtual Labs and Workflows

Virtual Laboratories (VLs) provide the tools, methods and infrastructure to enable e-science experimentation. Workflow management systems are key components of VLs, exposing the structure of experiments and enhancing interactive experimentation. We study workflows running on a grid-based VL in the context of biology research.

Ridges and significant workflows

A transciptome map is a map of the transcription activity of genes with respect to their chromosomal location. Such a map provides a global overview of the transcription activity of genes in a cell. This information may give some insight on how gene regulation works. Previous work analyzed a human transcriptome map generated using SAGE data and found regions of increased (median) gene expression (RIDGEs)[1]. On average, genes in ridges have a higher expression than genes outside these regions, irrespective of the analyzed tissue.
We extended the concept of RIDGEs so that they can be computed for any ordered sequence of values, including all kinds of genomic profiles, or even time series of temperature. In the general case, RIDGEs are called significant windows.

SigWin-detector workflow

We have developed SigWin-detector, a workflow that identifies significant windows in a given ordered sequence according to the method used in Versteeg et al. to identify RIDGEs [1]. The workflow is composed of modules that compute sliding window medians of the input sequence and identify significant windows, i,.e., windows with a median value above a certain false discovery rate (FDR) threshold (Figure 1). Final and intermediate results are visualized within the workflow environment. These visualizations enable interactive adjustment of the parameters of each module to explore the solutions space.

Figure 1 Schematic representation of the SigWin-detector workflow

Methods

Creating and running a workflow with VLAM-G

The VLAM workflow management system enables the creation and execution of workflows in a Grid environment [2]. With VLAM we can create and modify workflows by dragging and dropping (or deleting) the necessary modules into the workflow composer window and linking them with each other (see Figure 2). We use the experiment menu (Figure 3) to run the workflow interactively. In the experiment menu window it is possible to change the input parameters for each module and to visualize the results.
It is also possible to run an experiment in batch mode by calling the run time system of VLAM using a script. This is useful when the experiment will take a long time to run.

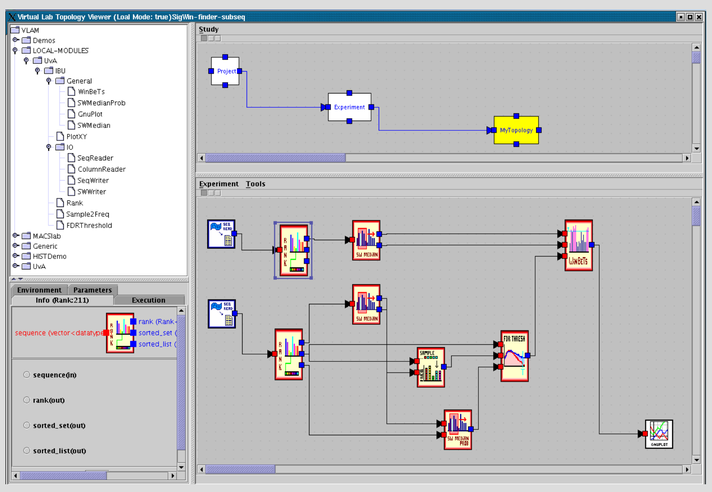

Figure 2

VLAM workflow composer displaying SigWin-detector workflow.

Computing false discovery rate thresholds

We developed a new way of computing the null hypothesis distribution needed by the FDR thresholds. This new method avoids the expensive step of computing sliding medians of permuted input data, thereby improving radically the overall performance.

Results

Identifying ridges

To validate our workflow, we used SigWin-detector to identify ridges in the human transcriptome map compiled by Versteeg et al. [1]. Our results partially shown in Figure 3 agree with their results.

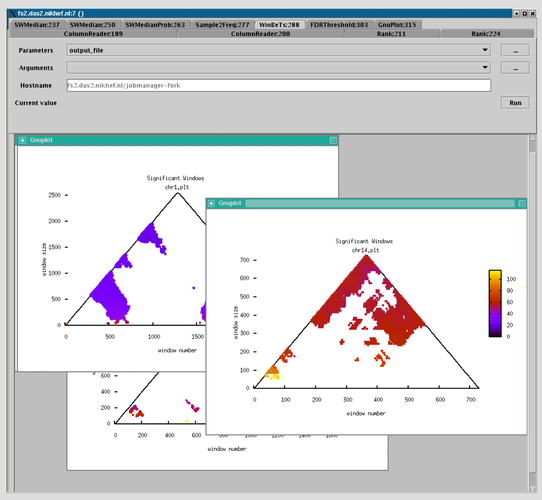

Figure 3

Ridges of the HTM for various chromosomes obtained using SigWin-detector.

Identifying summers in Amsterdam

We also tested our workflow using a time series of temperatures in Amsterdam. The resulting 'checker-board' pattern (Figure 4) clearly identifies the periodicity of the data.

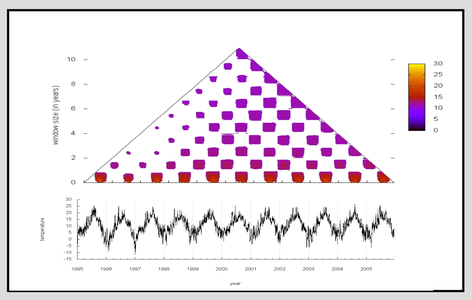

Figure 4

Temperatures in Amsterdam. The significant window regions correspond to the warmer months of the year.

Conclusions

- The SigWin-detector workflow identifies regions of unexpectedly high values in any kind of ordered sequence. It also identifies patterns such as periodicity.
- Our workflow based method generalizes and outperforms the original method, and it can be easily adapted to meet the user's specific needs.
- VLAM gives easy access to grid resources and can run workflows interactively and in batch mode.
- The development of SigWin-detector pointed out some weak points in the user interface of VLAM. Therefore, a new user interface is being developed.
- New features will include the possibility of using VLAM as a web service and interoperability with other workflow management systems such as Kepler and Taverna.

References

[2] R. Versteeg, B.D.C. van Schaik, et al. (2003). "The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes." Genome Research 13(9): 1998.
[3] H. Afsarmanesh, R.G. Belleman, A.S.Z. Belloum, et al. (2002). "VLAM-G: A Grid-based virtual laboratory." Scientific Programming 10(2): 173.

This work was part of the BioRange program of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

M.A. Inda *et al.*
Integrative Bioinformatics Unit, Kruislaan 318, Building I, 1098 SM, Amsterdam, The Netherlands

Email: **inda@science.uva.nl**
URL: **http://www.vl-e.nl/**
http://integrativebioinformatics.nl

This work was carried out in the context of the Virtual Laboratory for e-Science project. This project is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science (OC&W) and is part of the ICT innovation program of the Ministry of Economic Affairs (EZ).