Poster L-26

SigWin-finder: an interactive workflow in a virtual laboratory for e-bioscience

Authors:
Marcia A. Inda (*Integrative Bioinformatics Unit, Faculty of Science, University of Amsterdam*)
Marco Roos (*Integrative Bioinformatics Unit, Faculty of Science, University of Amsterdam*)
Adam S. Z. Belloum (*Institute of Informatics, Faculty of Science, University of Amsterdam*)
Dmitry Vasunin (*Institute of Informatics, Faculty of Science, University of Amsterdam*)
Timo M. Breit (*Integrative Bioinformatics Unit, Faculty of Science, University of Amsterdam*)

Short Abstract: Running on a grid-based virtual laboratory, the SigWin-finder workflow analyzes any given sequence of values, spanning from gene expression data to time series of temperature. It implements a generalized and improved version of the ridgeogrammer method to identify regions of increased gene expression in trancriptome maps.

Long Abstract:

MOTIVATION

We study workflows running on a grid-based enhanced virtual laboratory (VL-e) in the context of biology research. Virtual laboratories provide tools, methods and infrastructure to enable e-science experimentation in domain specific problem solving environments (PSE) [1]. Workflow engines are key components of PSEs, exposing the structure of experiments and enhancing interactive experimentation.

A transcriptome map is a profile of the transcription activity of genes in relation to their chromosomal location. Such a profile provides a global overview of the transcription activity of genes in a cell and may give some insight on how gene regulation works. In this work we present a generalized and improved version of the ridgeogrammer method to identify regions of increased gene expression (ridges) in transcriptome maps [2].

RESULTS

We have developed SigWin-finder, a general workflow that identifies significant windows in a given sequence. Significant windows are an extension of the concept of ridges for the case of an arbitrary input sequence. In the general context of significant windows, one can analyze different types of genomic profiles (and not only transcriptome maps), or even unrelated sequences such as time series of the temperature in Amsterdam.

The SigWin-finder workflow was implemented using VL-e VLAM-G toolkit (VLAM), which enables the interactive creation and execution of workflows in a grid environment [3]. VLAM exposes the modular structure of workflows permiting easy adaptation of SigWin-finder to meet the user specific needs. VLAM also offers the opportunity of running an experiment in batch mode by calling its run time system from a script (and shortly from a web service). This is a useful feature when the experiment takes a long time to run.

SigWin-finder workflow computes sliding window medians of the input sequence, identifying as significant the windows with median value above a certain false discovery rate (FDR)

threshold. Visualization modules display graphics of final and intermediate results within the workflow environment, enabling interactive adjustment of the parameters of each module to explore the solution space.

Our implementation uses a new method for computing the null hypothesis distribution needed to compute the FDR thresholds. Our method avoids the expensive step of computing sliding medians of permuted input data. As a result, SigWin-finder is orders of magnitude faster than its antecessor, doing in a few minutes what used to take many hours. This permits interactive use of SigWin-finder allowing the analysis of many transcriptome maps in a day. It also makes it possible to investigate much larger profiles, (e.g., DNA sequence based profiles). Furthermore, it is easy to parallelize the workflow using job farming techniques, allowing us to take maximum advantage of the grid resources available.

We used SigWin-finder to identify ridges in the human transcriptome map compiled by Versteeg and coworkers [2]. Our results agree with their results. We also tested our workflow using periodic sequences, e.g., time series of temperature in Amsterdam. The resulting checkers board pattern of significant windows clearly reveals the periodicity of the data, identifying the periods of temperature higher than the median (i.e., summers). At the moment we are running our workflow using different kinds of input data, including DNA sequence profiles, microarray data, and weather related data. We also intend to run it using different types of fabricated data to create a collection of patterns so that we can identify such signatures when they appear in an experimental input sequence.

REFERENCES

[1] Rauwerda H, Roos M, Hertzberger BO, Breit TM (2006) The promise of a virtual lab in drug discovery. Drug Discovery Today 11(5-6): 228-236.

[2] Versteeg R, van Schaik BD, van Batenburg MF, Roos M, 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): 1998-2004.

[3] Belloum ASZ, Groep DL, Hendrikse ZW, Hertzberger BLO, et al. (2003). VLAM-G: a Grid-based virtual Laboratory. Future Generation Computer Systems 19(2): 209-217.