Cloud-based bioinformatics with Apache Spark

Who can attend

Participants are expected to have at least some minimal prior experience in programming, using any language (R, Python, or Perl would suffice). A basic working knowledge of molecular biology is helpful but not essential.

Registration fees

JNU M. A. / M. Sc. Students : Free
JNU M.Phil/Ph.D students : ₹ **1000.00**
JNU Faculty : ₹ **2000.00**
Students of other recognised institutions : ₹ **2000.00**
Faculty of other recognised institutions : ₹ **4000.00**
Industry, private institutes : ₹ **10000.00**
Participants from outside India : ₹ **10000.00**

Maximum intake

A maximum of 50 participants will be accommodated. Selection (in case of excess requests) will be made by the course coordinator and teaching faculty based on the compatibility and usefulness of the course and also considering regional, social and gender diversity.

Registration

http://www.gian.iitkgp.ac.in

Last date for applying is **August 14, 2018**

Course Coordinator

Professor Shandar Ahmad
School of Computational and Integrative Sciences,
Jawaharlal Nehru University, New Delhi-110067
Phone: +91-11-2674-8788 (O)
Email: shandar@jnu.ac.in

GIAN, JNU

GIAN Global Initiative of Academic Networks

Cloud-based bioinformatics with Apache Spark

October 3-7, 2018
Jawaharlal Nehru University
New Delhi

SCHOOL OF COMPUTATIONAL & INTEGRATIVE SCIENCES (SC&IS)
Jawaharlal Nehru University, New Delhi

Overview

Cloud-based bioinformatics with Apache Spark

The computational needs of bioinformatics are constantly increasing. Although sophisticated ready-made tools are increasingly available, in order to fully control their methods, bioinformaticians and other data scientists will need to write or modify their own software. Recently, there has also been a shift in computational architectures, from single-core desktop and laptop computers to multicore and distributed systems such as cloud computing. This shift necessitates a change in the way that we approach programming and think about algorithms in general and also specifically in bioinformatics. Approaches such as MapReduce have become very popular for performing cloud based computational tasks. However, MapReduce is not always applicable to a given problem, because of the constraints of its computational model, and better solutions are becoming available.

In this course we will introduce Apache Spark (http://spark.apache.org), a framework that builds on the ideas of the MapReduce paradigm, but overcomes some of its limitations, to support practical cloud computing. A cutting-edge programming environment is also essential to utilize the full power of this sophisticated framework.

Thus, we will also introduce the Scala programming language, a modern and convenient language, which builds on recent advances in computer science and has seen rapid adoption in both the academic and industrial communities. We will develop tools to analyse the well-known Open TG-GATEs and similar toxicogenomics datasets (http://toxico.nibiohn.go.jp), carrying out this analysis using the Google Cloud Platform (http://cloud.google.com), which supports highly scalable computation, storage and deployment. Using these tools, we will study the toxicity effects of well-known drugs on metabolic pathways and individual genes.

Objectives

This course will give a theoretical background as well as hands-on experience in the following topics.

- Scalable computation with Spark on the Google Cloud Platform with Dataproc
- High-performance, concurrent algorithms for data analysis using the Scala language
- Software engineering for effective bioinformatics
- Toxicological data investigation using the Open TG-GATEs toxicogenomics dataset

Although our examples will focus on toxicology, the skills taught will be practically useful in a variety of scientific settings and for a wide range of computational problems.

Course details

1. Fundamental Scala programming
Basic functional programming. The benefits of strong typing. Collection operations. Compiling and running code. The interactive Scala console. Using Eclipse and SBT. Thinking about parallel computation.

2. The Open TG-GATEs dataset
Basic toxicogenomics. Working with samples. Analysing gene sets by enrichment testing. Additional Scala concepts. Reasoning about performance.

3. Using Spark on Google Cloud
Creating a cluster and deploying code. Running a parallel analysis on Google Dataproc. Accessing storage buckets. The interactive Spark console.

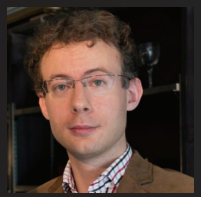
4. Toxicological data analysis with Spark (1)
Computing p-values, discovering genes and samples of interest. Resilient distributed datasets (RDDs), DataFrames and DataSets.

5. Toxicological data analysis with Spark (2)
Ranking samples by similarity. Validating a biological hypothesis. Reasoning about metabolic pathways.

Teaching and learning methods

Most of the course will be delivered in an interactive manner in a laboratory setting. Practical exercises and hands-on skill development will be the primary approach.

Teaching faculty



Dr Johan Nyström-Persson is a researcher and engineer at Douglas Connect GmbH, Switzerland, and Level Five, Japan. He has worked extensively with analysis of toxicological and NGS data. He is the main developer of the Toxygates application for toxicogenomics (NIBIOHN, Japan), and worked with professor Rudi Appels' team as part of the IWGSC effort to assemble chromosome 7A of common wheat (Murdoch University/ACCWI, Australia). His background prior to bioinformatics is in computer science, where he did research on virtual machines and programming languages (PhD University of Tokyo 2012). He was also formerly a software engineer at Google, and obtained his BSc from Imperial College London (2006). He is currently based in Tokyo.