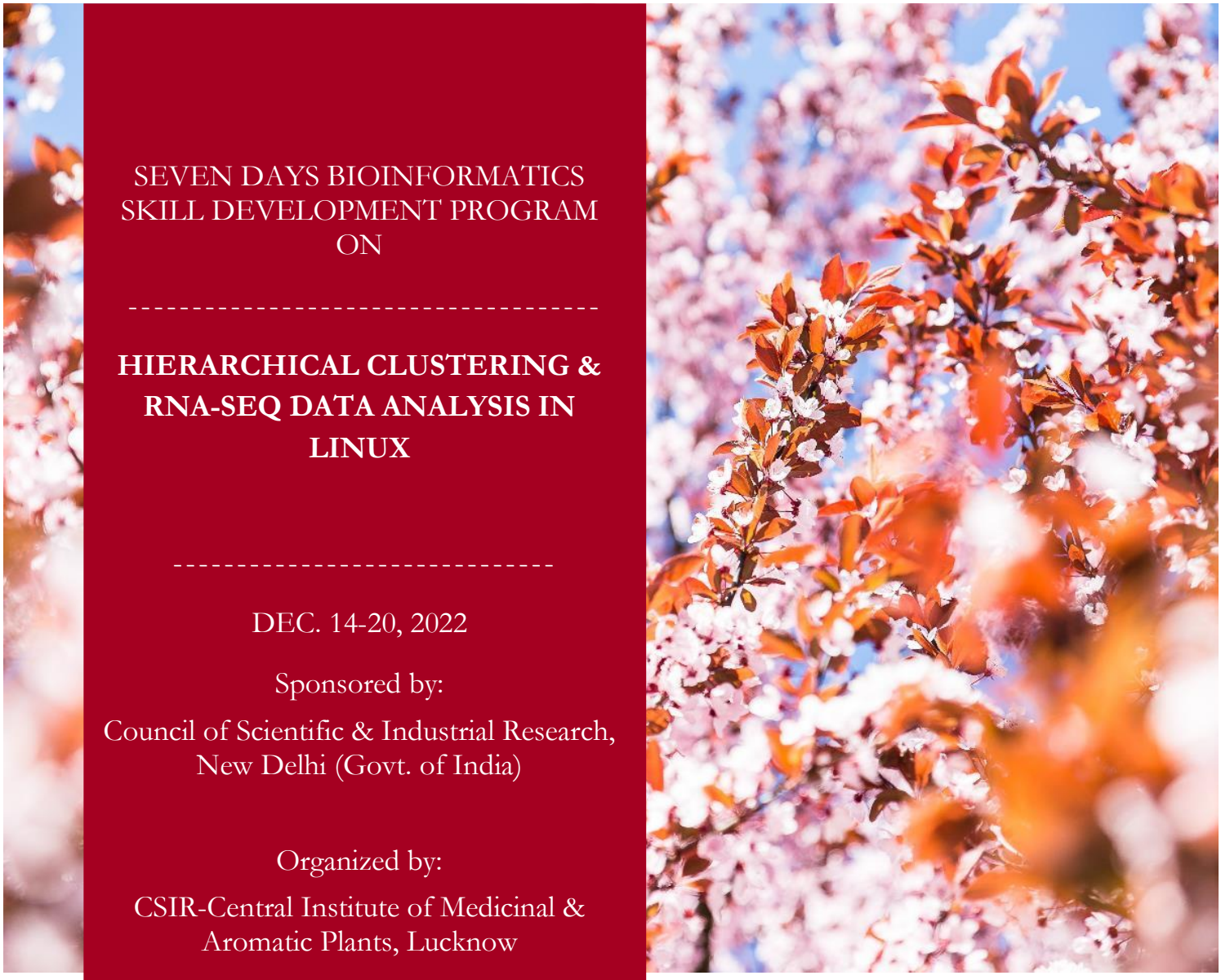

SEVEN DAYS BIOINFORMATICS SKILL DEVELOPMENT PROGRAM ON

HIERARCHICAL CLUSTERING & RNA-SEQ DATA ANALYSIS IN LINUX

DEC. 14-20, 2022

Sponsored by:

Council of Scientific & Industrial Research, New Delhi (Govt. of India)

Organized by:

CSIR-Central Institute of Medicinal & Aromatic Plants, Lucknow

Training Program Contents

- Transcriptome data retrieval (SRA/ENA)
- Data quality check (FastQC)
- Trimming of garbage sequences (Cutadapt)
- *De-novo* of high-quality reads (TRINITY)
- Assessment of assembled reads (BUSCO, N50)
- Non-redundant data clustering using hierarchical clustering (CD-HIT)
- Gene expression quantification (RSEM)
- Differentially expressed gene (DEG) analysis (EdgeR)
- BLAST, Pathway, and Gene Ontology enrichment analysis.
- R programming and plots generation (Heatmap and Volcano plot)

Convenor
Dr. Feroz Khan

Coordinator
Dr. Laiq-ur Rahman

Chairman
Dr. P.K. Trivedi
Director, CSIR-CIMAP
Lucknow

About CSIR-CIMAP, Lucknow

CSIR-Central Institute of Medicinal & Aromatic Plants (CSIR-CIMAP) is a premiear multidisciplinary research institute of Council of Scientific & Industrial Research (CSIR), New Delhi, India with its major focus on exploiting the potential of medicinal and aromatic plants (MAPs) by cultivation, bioprospection, chemical characterization, extraction, and formulation of bioactive phytomolecules. With a strenght of 100 scientists, 162 technical officers, 129 support staff and nearly 300 doctoral and post-doc scholars at its head-quarter in Lucknow and research centers at Bengalure, Hyderabad, Pantnagar, and Purara. CSIR-CIMAP has played a key role in positioning India as a global leader in production of mints, vetiver and other aromatic grasses, and in ensuring indigenous production of artemisinin – a WHO approved antimalarial. CSIR-CIMAP houses a National Gene Bank on MAPs, which is one of the three of its kind in India. CSIR-CIMAP has played a key role in successfully commercializing an ayurvedic herbs based antidiabetic formulation, which has now benefitted millions. The institute is presently accredited by ICS-UNIDO and Indian-Ocean Rim Association (IORA) as a focal point for research and training on Medicinal Plants among 21 participating member countries. For more details please see the CSIR-CIMAP website www.cimap.res.in

About Bioinformatics Skill Development Program

'Omic' technologies cover universal detection of genes (genomics), mRNA (transcriptomics), proteins (proteomics) and metabolites (metabolomics) in a specific biosample. Data analysis is complicated as a massive amount of data generated, and bioinformatician involvement in the process is essential. Transcriptomics data mining is an efficient way to discover genes or gene families encoding enzymes involved in various metabolic pathways. High-throughput next-generation sequencing (NGS) technologies have revolutionized transcriptomics especially with the advent of RNA-sequencing (RNA-seq). This technology can be used to obtain RNA sequences on a massive scale with enormous sequencing depth. Plants produce a vast array of specialized metabolites, many of which used as pharmaceuticals, flavors, fragrances, and other high-value fine chemicals. Most of these compounds occur in non-model plants for which genomic sequence information is not yet available. The production of a large amount of nucleotide sequence data using next-generation technologies is now relatively fast and cost-effective, especially when using the latest Roche-454 and Illumina sequencers with enhanced base-calling accuracy. To investigate specialized metabolite biosynthesis in plants establishment of data-mining framework required by employing next-generation sequencing and computational algorithms, to construct and analyze the transcriptomes of plants that produce compounds of interest for biotechnological applications. After sequence assembly an extensive annotation approach required to assign functional information to transcripts. The annotation based on direct searches against public databases, e.g., RefSeq, InterPro, GO, EC, and associated KEGG pathway maps. This study aims to identify biosynthetic gene candidates related to specific metabolic pathways. These assembled transcriptome data access through web-based BLAST server. Transcriptomes are studied for interpreting functional elements of the genome and revealing molecular constituents of cells and tissues.

The Aim of Bioinformatics Skill Development Training Program

To familiarize students/researchers/academicians/industry experts with the basics of machine learning method e.g., Clustering, hierarchical clustering and its use in RNA-Seq data analysis especially Heat-map/Dendogram tree representation of Differentially Expressed Genes (DEGs). Participants may understand the role of Hierarchical Clustering in Dendogram tree generation and interpretation of Heat-map. In parallel, practical exercises/example demo for technical skill development will be scheduled after introductory lectures. Participants need to follow the instructions and perform the different steps during Hands-On training. Live trouble shooting will assist the participants in smooth learning of tools and techniques. The training program will cover an invited expert lecture, a training program theme lecture and demo presentation/practical exercise session. The training program would cover the following aspects:

- Installation and setup of required software and packages e.g. TRINITY and their associated packages on Linux OS.
- Transcriptome data fetching
    - Sequence Read Archive (SRA)
    - European Nucleotide Archive (ENA)
- Quality check of selected transcriptome data using FastQC software.
- Pre-processing of raw read files (FASTQ) in order to remove low-quality reads, noise sequences, etc, using the Cutadapt tool.
- *De-novo* transcriptome assembly of high-quality reads using TRINITY software.
- Assessment of assembled high-quality reads
    - Benchmarking Universal Single-Copy Orthologs (BUSCO)
    - N50 (Trinity stats)
    - Total alignment rate (Bowtie2)
- Hierarchical clustering of high-quality assembled reads using CD-HIT pipeline, to generate non-redundant reference transcripts.
- Introduction to Hierarchical Clustering and its application in Dendogram tree generation.
- RSEM pipeline for abundance estimation based on the mapping of RNA-seq reads to TRINITY assembled contigs.
- The differential gene expression (DEG) analysis of selected samples using the EdgeR (Bioconductor package).
- Functional annotation using Standalone BLAST+ pipeline against UniProt database.
- Pathway mapping (KEGG) and gene ontology (GO) enrichment analysis.
- Visualization of DEG result by making Hierarchical clustering Heatmap and Volcano plot.

The skill development training gained through this program may help in making the career in Biotechnology, Bioinformatics, Functional Genomics, Machine Learning, Big Data and Data Science.

Eligibility

UG/PG Science/Engg./Pharmacy students (Bioinformatics/ Biotechnology/ Pharmacy/ Biochemistry/ Microbiology/ Life Sciences/ Chemistry/ Botany/ Zoology/ Plant Sciences/ BioMedical Sciences; Ph.D. fellows/ Post-Doc scholars/ RA/ Scientist fellows/ Technical Officers; Project fellows/ Industry Professionals/ Entrepreneurs/ Academicians/ Faculty can attend. Basic knowledge of Biology, Chemistry, Mathematics, Statistics, and working experience of Windows OS and Linux OS (Ubuntu) is required. For fresher's, tutorial will be provided for basics commands of Linux.

Certification

Training program's participants will receive a digital certificate of participation from the CSIR-CIMAP, Lucknow after successful completion of the skill development program. The digital certificate will be emailed after the successful completion of training program.

Feedback

After training program, participants may be asked to submit the given feed-back form. Participants may be asked to express their training experiences and suggestions for further improvement.

Technical requirements

Participants may have Laptop/Desktop PC with Windows 10 OS along with Linux Ubuntu sub-system app (installed it from Microsoft Store), Virtual Machine with Linux OS, knowledge of DOS commands, Scripting knowledge on any text editor e.g., Notepad, Notepad++, Vi editor and working experience of MS Office and internet browsing. A working version of Windows OS and MS Office software are necessary to follow the practical examples/training sessions online. Knowledge of basic cell biology and biostatistics will be beneficial. Compatible hardware includes machine with 16-32 GB RAM and 2-4 GB Graphics Card, Intel's/AMD 6-8 Cores CPUs and medium series motherboard with heat sinks, and 1-2 GB HDD/SSD/NVMe storage.

Training mode: Offline & Online hybrid mode. Online mode training will be done through MS Team/Google Meet or similar online apps.

Registration Fee: Rs.5,000/- for each participants. **Rs.10,000/-** for Industry Professionals

The registration fee includes digital Registration kit which includes training brochure, program schedule, tutorials, invited/expert lectures, practical exercises, feed-back form and a certificate (digital copy will be emailed, however participants may collect the hard copy or request for speed-post) after successful completion of the training.

Registration fee can be pay through online mode to SBI bank A/c No. 00000030267691783, SBI Main branch, Hazratganj, Lucknow (IFSC code: SBIN0000125) or through Demand Draft in favor of '**Director, CIMAP**', payable to Lucknow. Complete registration form along with the fee details should reach us (f.khan@cimap.res.in; l.rahman@cimap.res.in) before deadline i.e., Dec. 13, 2022 upto 5:00 PM. Registration to the skill development training program will be on 'First-come-First-serve' basis. Seats are limited.

For any query related to this skill development training program, kindly contact:

Dr. Feroz Khan, Convenor (f.khan@cimap.res.in) Mob. 9415538701/ Ph.(O) +91 522 2718668

Dr. Laiq-Ur Rahman, Coordinator (l.rahman@cimap.res.in)

For any further details please contact:

Dr. P.K. Trivedi,

Director

CSIR-Central Institute of Medicinal & Aromatic Plants,

P.O.-CIMAP, Kukrail Picnic Spot Road, Lucknow-226015, INDIA

Ph.: +91 522 2718639, 2718641, 2718505

E-mail: director@cimap.res.in , Website: www.cimap.res.in

Application/Registration Form

Candidate Full Name: _____

Designation/Position: _____

Affiliation (Institute/Univ.):_____

Address:_____

Locality Type (Urban/Rural):_____

Category (Gen/OBC/SC/ST):_____

Gender (Male/Female):_____

Area of Interest: _____

E-mail: _____

Contact No.: (+91) _____

Payment Details:

Registration Fee: Rs. _____

Mode of Payment (Online/UPI/DD): _____

Online/UPI Transaction/DD No._____ Date _____

Bank Name: _____

Name _____ Signature_____