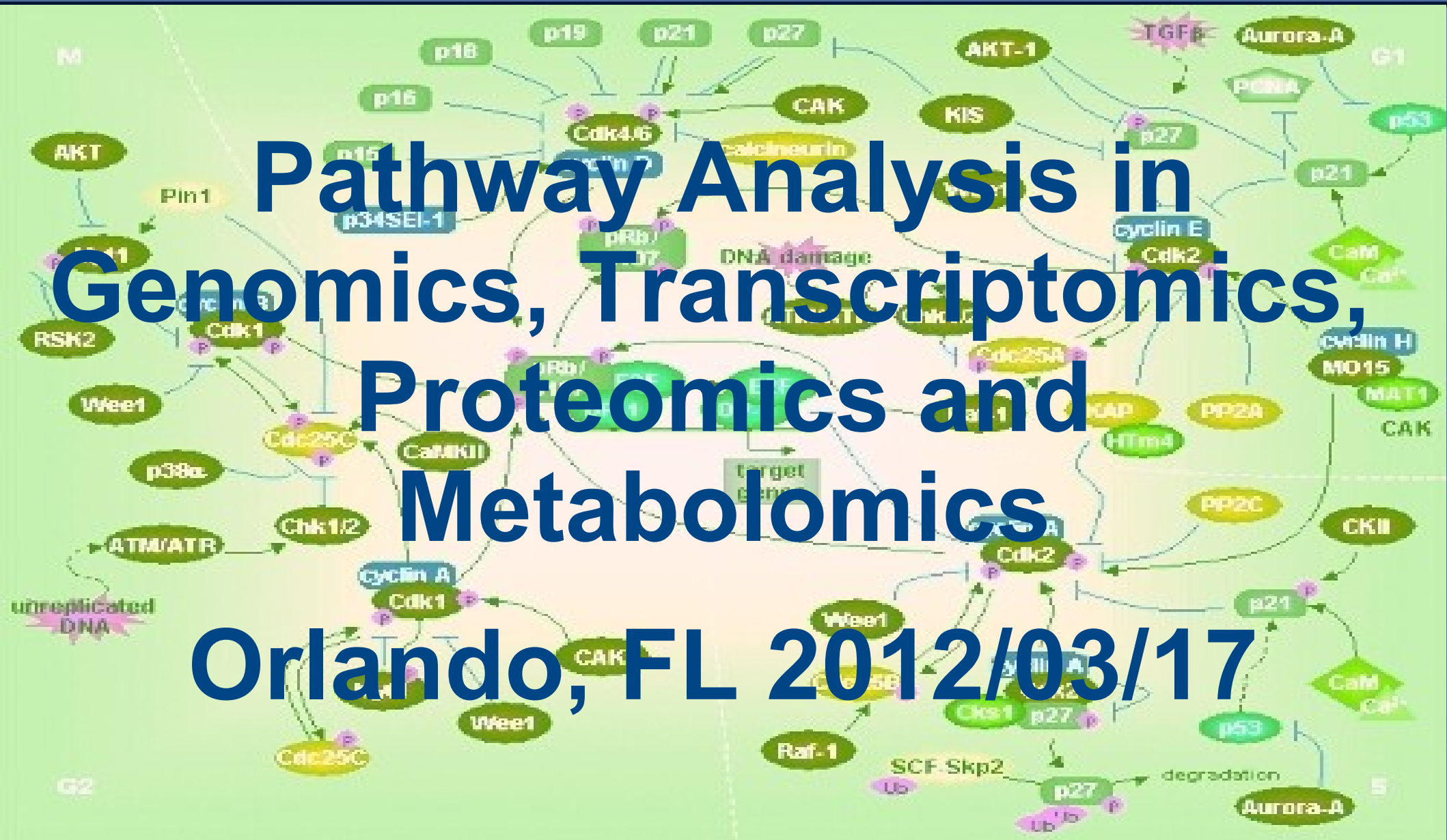
ABRF Workshop



Pathway Analysis in Genomics, Transcriptomics, Proteomics and Metabolomics

Orlando, FL 2012/03/17

Why PA now?

Driven by the need to understand mechanisms in Biology

Fed by a torrent of “omics” data

Integrated in Systems Biology approaches

Towards P4 Medicine, environmental, pharma

What is PA

A set of techniques that provide insight into the structure and functioning of biological systems, allowing for inference on their behavior.

The methods start from large lists of genes and their annotation, use quantitative levels of expression and produce reduced lists and/or models.

Training issues

- Three day long training courses in Oeiras, PT
- For about 20 people
- Connected to other themes:
  - Gene Regulatory Networks (2009)
  - Chemoinformatics (2010)
  - Network Biology (2011)
- Results: need more time, more statistics bkgnd

Well conducted PA will assume

Verification of the assumptions

Statistics proficiency (trained users)

High quality of the annotations

We need more and better knowledge bases

Abundance of omics data

More (open) data from tech services, targeted research

Basics

- From a list of genes from HT experiments in two different conditions, we want to be able to reliably extract reduced lists of genes that justify the functional change.

- We also want to interpret the above findings in the context of biological mechanisms being present/absent.

Knowledge bases

Provide annotated gene sets

KEGG

REACTOME

METACYC

GO

Types of PA tools

ORA (Over-Representation Analysis) ex: FatiGo

FCS (Functional Class Scoring) ex: segPathway

PT (Pathway Topology) ex : GenExplain

Reference

Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges

Purvesh Khatri, Marina Sirota, Atul J. Butte

PLOS Computational Biology, Feb 2012, vol. 8, Issue 2, e1002375

ORA Methods

- Works on counts of genes
- Ignores signal “strength”
- Filters-out by thresholding
- Assumes independence between pathways

FCS approaches

- Statistical classification and grouping
- Correlation with phenotypic characters
- Significance test between Pw statistics

PT-based techniques

Like FCS, but make good use of co-expression and interaction annotations

Single-out topological models that explain the data

All the above

May produce defective results

In presence weak annotation quality

Failure to meet statistical assumptions

Artifacts

However...

We need to be ready for upcoming data deluges and their consequences

Medical information coming from new corners

- adverse effects database (Stanford)

- patients data collections

New developments

We need

- a significant investment in software development
- data interoperability and standards for cross-platform exchange
- better validation strategies

We need trained users