Aalto University
School of Science

HELSINKI
INSTITUTE FOR
INFORMATION
TECHNOLOGY

UNIVERSITY OF HELSINKI

Targeted Retrieval of Gene Expression Measurements Using Regulatory Models

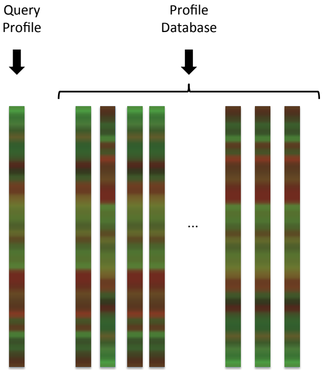
Elisabeth Georgii, Jarkko Salojärvi, Mikael Brosché,
Jaakko Kangasjärvi, Samuel Kaski

Funding: Tekes, Academy of Finland, PASCAL2

MLSB 2012

Motivation

- Large repositories of measurement data $\Longrightarrow$ use them!
- Goal: automated search for relevant experiments
- Considered task: given a gene expression profile, find “similar” profiles from a database



What is a suitable similarity measure?

- Shared keywords in the annotation (= knowledge-driven)
  (+) reliable, state of the art; (-) excludes new findings

  (Zhu *et al.*, Bioinformatics, 2008)

- Correlation of profiles (= data-driven)
  (+) easy to compute; (-) ignores gene dependencies

  (Engreitz *et al.*, BMC Bioinformatics, 2010)

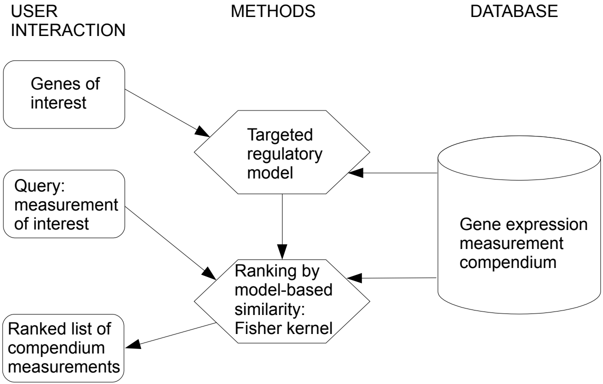
- Model-based similarity measure (= data-driven)
  (+) learns from database; (-) computationally expensive

  (Caldas *et al.*, Bioinformatics, 2009, 2012)

This approach: Model-based targeted retrieval

- Two main aspects
- **Targeted focus:** guide the model by genes of interest
  e.g. genes known to be related to a certain disease
  → adapt to users' needs, reduce computational effort
- **Similarity based on gene regulatory network models:**
  potential similarity of conditions at detailed biological level
  → improved interpretability by network activation patterns

System for targeted retrieval


USER INTERACTION

METHODS

DATABASE

Genes of interest

Query: measurement of interest

Ranked list of compendium measurements

Targeted regulatory model

Ranking by model-based similarity: Fisher kernel

Gene expression measurement compendium

- First step: learn regulatory model for user-provided genes
- Second step: retrieve measurements related to a query

Targeted gene expression model

- Conditional model: expression of target genes, given expression of other genes

$$P(X_{\mathcal{T}} | X_{-\mathcal{T}})$$

- Pseudo-likelihood approach:

$$\tilde{P}(X_{\mathcal{T}} | X_{-\mathcal{T}}) = \prod_{j \in \mathcal{T}} P(X_j | X_{-\{j\}}; \theta_j)$$

  i.e., independent model for each target gene

- Gene-specific model: Gaussian linear regression model

$$X_j = X_{-\{j\}} \beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

  sparse β estimate by L_1-norm regularization
  $\rightarrow$ target gene neighbors

Model-based similarity measure

- Fisher score representation of data point: $s_{\hat{\theta}}(x^{(i)})$:
  gradient of its log-likelihood at learned model parameters
  $\rightarrow$ direction in which to update the parameters after adding $x^{(i)}$ to the dataset ($\rightarrow$ summary of dataset $D + x^{(i)}$)

- Simple Fisher kernel: (Jaakkola and Haussler, NIPS 1998: using HMMs in classifiers)

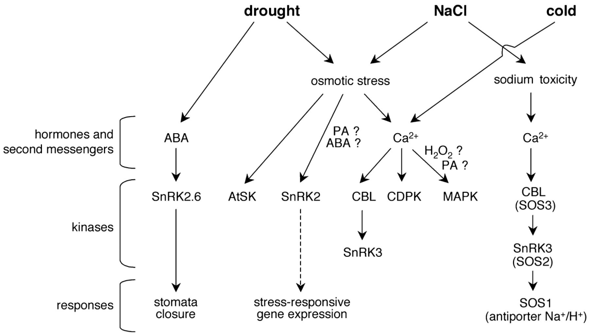
  $$K_{\hat{\theta}}(x^{(i_1)}, x^{(i_2)}) = s_{\hat{\theta}}(x^{(i_1)})^T s_{\hat{\theta}}(x^{(i_2)})$$

  $\rightarrow$ similarity of datasets $D + x^{(i_1)}$ and $D + x^{(i_2)}$ regarding model-based summary statistics

- Parameters of biological interest in our model:
  coefficients of target gene neighbors

Case study on plant osmotic stress

- Osmotic stress: dehydration of plant
- Causes: drought, salt, or cold conditions
- Relevance: important abiotic stress for crop productivity
- Cellular response:



Boudsocq M , Laurière C Plant Physiol. 2005;138:1185-1194

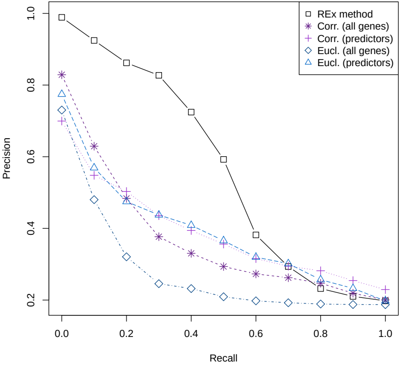
Case study on plant stress

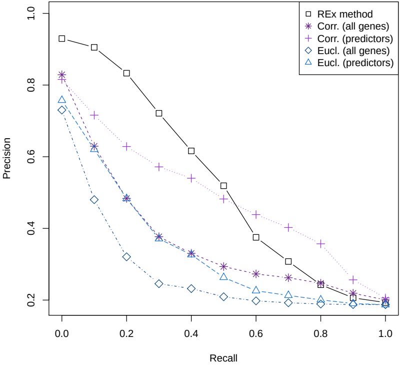
- Data: 141 differential expression profiles from 38 *A. thaliana* stress datasets, 6658 diff. expr. genes
- Task: retrieval of osmotic stress experiments (31 profiles from 5 datasets, $\geq$ 6 profiles per dataset)
- Target gene lists from two sources:
  - 10 water-stress related genes (TF DREB2A + targets)

    (Sakuma *et al.*, PNAS, 2006)
  - 41 genes annotated as 'drought-salt-cold'

    (STIFDB, Shameer *et al.*, Int J Plant Genomics, 2009)
  - *overlap: 4 genes*
- Experimental setup:
  - One left-out dataset as queries (cross-validation)
  - Unsupervised model training with all other profiles (including osmotic and non-osmotic)

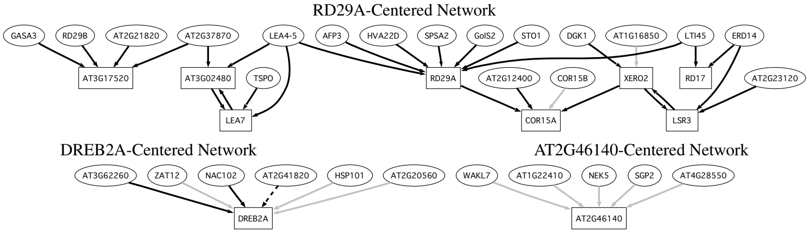
Precision-recall analysis


Target list: Sakuma-water


Target list: STIFDB

- Modeling targeted gene relationships helps

Osmotic stress network analysis



RD29A-Centered Network

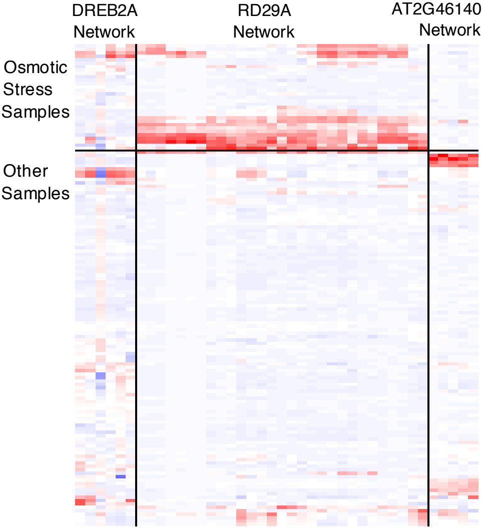
DREB2A-Centered Network

AT2G46140-Centered Network

- Top edges in bootstrapping

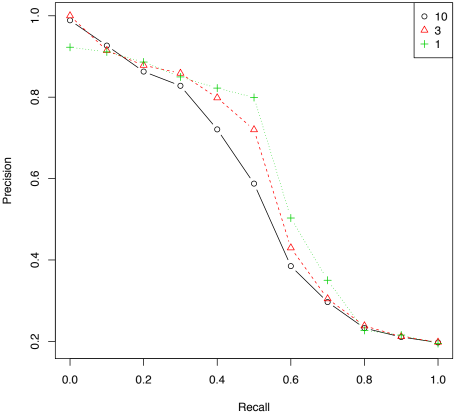
Target	Predictor	Stress-related annotation of predictor?
RD17	LTI45	yes (also included in STIFDB)
COR15A	COR15B	yes (also included in STIFDB)
XERO2	LSR3	yes (also included in Sakuma-water)
RD29A	LTI45	yes (also included in STIFDB)
AT3G02480	LEA7	yes (also included in Sakuma-water)
AT1G52690	AT3G02480	yes (also included in Sakuma-water)
LSR3	XERO2	yes (also included in Sakuma-water)
LSR3	ERD14	yes (also included in STIFDB)
AT3G17520	RD29B	yes (response to water deprivation)
RD17	ERD14	yes (also included in STIFDB)
DREB2A	AT3G62260	– (protein phosphatase 2C)
DREB2A	ZAT12	yes (involved in cold acclimation)

Model-based comparison of measurements



Discriminative target genes

- Test performance of optimal subsets of size *k*



- Best subset of size 1: RD29A (**r**esponsive to **d**ehydration)
- Best subset of size 3: RD29A, LEA7, COR15A

Discussion

- Summary: targeted retrieval using regulatory model
- Purpose: investigating specific commonalities between biological conditions based on (putative) gene relationships
- Efficiency: gene-specific models can be pre-computed
- Open questions:
  - Given promising performance with simple model, what is the most suitable model for retrieval? (also supervised options, prior knowledge, . . .)
  - Is the conceptual idea feasible for applications with heterogeneous data? (different platforms, species, measurement types, . . .)