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 \(\Rightarrow\) 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

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
  \[ P(X_T | X_{\cdot T}) \]

- Pseudo-likelihood approach:
  \[ \tilde{P}(X_T | X_{\cdot T}) = \prod_{j \in T} P(X_j | X_{\cdot \{j\}}; \theta_j) \]
  i.e., independent model for each target gene

- Gene-specific model: Gaussian linear regression model
  \[ X_j = X_{\cdot \{j\}} \beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \]
  sparse \( \beta \) 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 → 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)

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

→ 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**:

![Diagram of osmotic stress response](image_url)

- Hormones and second messengers:
  - ABA
  - SnRK2.6
  - AtSK
  - SnRK2
- Kinases:
  - CBL
  - CDPK
  - MAPK
- Responses:
  - Stomata closure
  - Stress-responsive gene expression

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

©2005 by American Society of Plant Biologists
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

<table>
<thead>
<tr>
<th>Target</th>
<th>Predictor</th>
<th>Stress-related annotation of predictor?</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD17</td>
<td>LT145</td>
<td>yes (also included in STIFDB)</td>
</tr>
<tr>
<td>COR15A</td>
<td>COR15B</td>
<td>yes (also included in STIFDB)</td>
</tr>
<tr>
<td>XERO2</td>
<td>LSR3</td>
<td>yes (also included in Sakuma-water)</td>
</tr>
<tr>
<td>RD29A</td>
<td>LT145</td>
<td>yes (also included in STIFDB)</td>
</tr>
<tr>
<td>AT3G02480</td>
<td>LEA7</td>
<td>yes (also included in Sakuma-water)</td>
</tr>
<tr>
<td>AT1G52690</td>
<td>AT3G02480</td>
<td>yes (also included in Sakuma-water)</td>
</tr>
<tr>
<td>LSR3</td>
<td>XERO2</td>
<td>yes (also included in Sakuma-water)</td>
</tr>
<tr>
<td>LSR3</td>
<td>ERD14</td>
<td>yes (also included in STIFDB)</td>
</tr>
<tr>
<td>AT3G17520</td>
<td>RD29B</td>
<td>yes (response to water deprivation)</td>
</tr>
<tr>
<td>RD17</td>
<td>ERD14</td>
<td>yes (also included in STIFDB)</td>
</tr>
<tr>
<td>DREB2A</td>
<td>AT3G62260</td>
<td>– (protein phosphatase 2C)</td>
</tr>
<tr>
<td>DREB2A</td>
<td>ZAT12</td>
<td>yes (involved in cold acclimation)</td>
</tr>
</tbody>
</table>
Model-based comparison of measurements

<table>
<thead>
<tr>
<th>Osmotic Stress Samples</th>
<th>DREB2A Network</th>
<th>RD29A Network</th>
<th>AT2G46140 Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other Samples</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Discriminative target genes

- Test performance of optimal subsets of size $k$

- Best subset of size 1: RD29A (responsive to dehydration)
- 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, . . . )