Clustering of Gene Expression Time Courses: Methods and Validity of Solutions

Ivan G. Costa Filho
Alexander Schliep

Computational Biology Department
Max-Planck-Institute for Molecular Genetics, Berlin
Goal
(1) Clustering of gene expression time-courses

(2) Validating clusterings of genes

(3) Methods for clustering heterogeneous data

gene expression + functional annotation, regulatory region information, protein-protein interactions ...
(1) Clustering Method

*Time-course models*
Biological Truism

• Many genes have
  – multiple functions
  – are involved in several regulatory networks
• *Unique assignment to groups dubious*
Method Outline

1. Define a class of statistical models for time-courses
2. Combine them in a mixture model
3. Decode the mixture to infer groups
Time-course models
Example: Up-regulation

Segment length

Location and width

expression

1.5
1
0.5
0
-0.5

1 2 3 4

time

Max Planck Molecular Genetics: Algorithmics Group
ghmm.org/gql
Prototype: Up-regulation

Location and width

Hidden Markov Model (HMM)

Segment length
Example: Constant

expression

0.5
1
1.5

0

-0.5

1 2 3 4

time
Prototype: Constant

Hidden Markov Model (HMM)
Viterbi Path: Up-regulation

expression

State

1 1 2 2

1 2 3 4
time
Viterbi Path: Up-regulation

State 1: 1 1 2 2
State 2: 1 2 2 2

Expression levels over time.
Prototype: Cyclic
Perspective

- \#states = \#time-points:
  - linear HMM " multi-variate Gaussian
  - covariance matrix diag(\(\sigma_1, \ldots, \sigma_t\))
- Typically \#states << \#time-points
Mixture of HMMs
Mixture models

• Mixture components: HMMs $\lambda_1, \lambda_2, \ldots, \lambda_k$
• Mixture model ” weighted sum of $\lambda_i$

$$P[\text{gene} | \text{mixture}] = \alpha_1 P[\text{gene} | \lambda_1] + \alpha_2 P[\text{gene} | \lambda_2] + \ldots$$

$\alpha_i \geq 0$, add to unity
HMM-based ‘Clustering’

- Input:
  - genes profiles $g_i$
  - collection of $k$ HMMs

- Initialization:
  - Assign the probability that a data-point belongs to each $k$ HMMs randomly

- Iteration (until convergence of assignment):
  - Compute the new HMM parameters (B-Welch)
  - Re-assign $g_i$ to a HMM proportionally to $P[gene | \lambda_\gamma]$
Inference of groups
From Mixtures to Groups

• Posterior probability of mixture component $\lambda_i$
  
  $$P[\lambda_i \mid gene]$$

• Shannon-entropy
  
  $$H( \{ P[\lambda_i \mid gene] \}_{1 \leq i \leq k} )$$

quantifies level of ambiguity in assignment
From Mixtures to Groups

- Choose entropy threshold $\tau$
- If entropy of posterior is below $\tau$
  - Assign gene to group $i$ of maximal posterior
- Else:
  - leave gene unassigned
Results
Diurnal Time Courses
Results

Diurnal Time Courses
(2) Cluster Validation
Motivation

- Which clustering method to use?
- How good is a clustering?
- GO enrichment gives no comparative basis!
Use **gene annotation (GO)** as external data to validate **mixtures** (or clusterings) from gene expression
In clustering, external indices look for the number of genes pairs that

<table>
<thead>
<tr>
<th>Same categ.</th>
<th>Distinct categ.</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>
External Indices

\[
corr. \text{Rand} = \frac{\#TP + \#TN - n_c}{\#Pairs - n_c}
\]
External Indices for Mixture Model

Given mixture models \( U \) and \( V \), consider the posteriors of the mixture components:

\[
\{ P[u_i \mid g] \}_{1 < i < C} \quad \{ P[v_j \mid g] \}_{1 < j < R}
\]

\( g_k \equiv g_l \leftrightarrow \) the event of co-occurrence of \( g_k \) and \( g_l \)

\[
P[g_k \equiv g_l \text{ given } U] = \sum_{j=1}^{C} P[u_j \mid g_k] \cdot P[u_j \mid g_l]
\]
External Indices for Mixture Model

\[ TP = \sum_{k=1}^{N} \sum_{l=k+1}^{N} P[g_k \equiv g_l \text{ given } U] \cdot P[g_k \equiv g_l \text{ given } V] \]

\[ TN = \sum_{k=1}^{N} \sum_{l=k+1}^{N} P[g_k \equiv g_l \text{ given } U]^C \cdot P[g_k \equiv g_l \text{ given } V]^C \]
Biological Data Experiments

- **Gene expression data:**
  - Yeast during sporulation (7 time points)
  - 1027 genes after 2 fold filtering
- **‘Clustering’ Methods:**
  - Hierarchical clustering (Pearson correlation)
  - K-means (Pearson correlation)
  - Mixture of HMMs
  - Mixture of Multivariate Normals (full covariance)
Biological Data Representation

- Mixture Model representation
  - Use each GO term as a component in the mixture
  - Maximum likelihood estimator of a multinomial distribution

\[
P[t_i \mid g] = \begin{cases} 
\frac{1}{\# \{j \mid g \in t_j\}}, & \text{if } g \in t_i \\
0, & \text{otherwise}
\end{cases}
\]
Results
Methods x ‘All’ GO Terms

The graph shows the performance of different methods (HMMs, Multi. Norm., K-means, Hierar.) across varying entropy thresholds. The ECR (possibly Error Corrected Rate or another metric) is plotted against the entropy threshold. The blue line representing HMMs increases sharply from 0.35 to above 0.65 as the entropy threshold decreases from 1.9 to 0.1. The green line for Multi. Norm. shows fluctuations around 0.35. The red line for K-means remains steady at 0.35, and the cyan line for Hierar. remains at 0.2.
(3) Clustering of Heterogeneous data
Motivation

Use additional large scale biological data to improve clustering of gene expression time-courses
Challenges of Heterogeneous Biological Data

Gene Ontology

Gene Expression

Location Analysis
Our Approach

• Semi-supervised learning
  – Encode location analysis as soft pairwise constraints
  – Mixture estimation with constraints (Lange et al., 2005, Lu and Leen, 2005)
Semi-Supervised Learning
Semi-Supervised Learning

Component #1          Component #2

positive constraints
Semi-Supervised Learning

- Component #1
- Component #2

Positive constraints
Negative constraints
Semi-Supervised Learning

Component #1                        Component #2

 positive constraints
              negative constraints

Max Planck Molecular Genetics: Algorithmics Group
Semi-Supervised Learning

In a semi-supervised learning context, the diagram illustrates two components, 

- **Component #1** represented by a red circle
- **Component #2** represented by an orange circle

The data points are scattered across two dimensions, $x_1$ and $x_2$, with negative constraints indicated by red lines and positive constraints by green dashed lines. The algorithm aims to classify new data points based on these constraints and existing labeled data.
Pairwise Constraints
Location Analysis

TF1 → g1
TF2 → g2
TF3 → g3
TF3 → g4

$w^+_{(1,2)} = 0.5$
$w^-_{(1,2)} = 0.5$
$w^+_{(2,3)} = 0.0$
$w^-_{(2,3)} = 1.0$
$w^+_{(2,3)} = 1.0$
$w^-_{(2,3)} = 0.0$
Mixture Estimation with Constraints (1)

Maximize the complete likelihood:

\[ P[X, Y|W, \Theta] = P[X|Y, \Theta] P[Y|W, \Theta] \]

where \( X \) is the observable data, \( Y \) the hidden data, \( \Theta \) the model parameters, \( W = \{W^+, W^-\} \) the pairwise constraints

The prior can be decomposed at:

\[ P[Y|\Theta, W] = P[Y|\Theta] P[W^+|Y, \Theta] P[W^-|Y, \Theta] \]
The posterior assignments are approximated by means of Gibbs sampling (Lu and Leen, 2005)
Data

• Gene expression data
  – time-courses of 384 genes during mitotic cell division in Yeast (Cho, 1998)
  – expert classification into ‘five’ cell-cycle phases

• Constraints
  – transcription factor location analysis (Lee, 2002)
  – true labels
Results
Constraints from True Labels

5% of gene pairs constrained
Constraints from Location Analysis

40% of gene pairs constrained
Possible Explanations

- Non-specific information content
- Noise in the data
- ...

Max Planck Molecular Genetics: Algorithmics Group
Constraints from True and Random Labels
Filtered Constraints from Location Data

Non-Filtered

Filtered

TR - Pos. and Neg. Constraints

0.454

TR Filtered - Pos. and Neg. Constraints

0.472
Summary

• Clustering of Time-Series
  – Flexible: cyclic & transient time-courses
  – Interactive & robust

• Cluster Validation
  – Methodology for evaluating clustering given functional annotation

• Heterogeneous Analysis
  – Successful integration of location analysis
Outlook

- Cluster Validation
  - Perform a ‘extensive’ evaluation of clustering methods
- Heterogeneous Analysis
  - learning of relevant constraints
  - *In-situ hybridization, protein-protein interactions*, …
- Clustering of Development Trees
  - In progress …
Acknowledgements

• A. Schönhuth and C. Steinhoff

• GHMM (http://ghmm.org):
  Wasinee Rungsarityotin, Benjamin Georgi, Ben Rich, Matthias Heinig,
  Alexander Riemer, Janne Grunau

• T. Beissbarth for help with GO
Thanks.

ghmm.org/gql
Motivational Buzzphrases

Model qualitatively

Embrace ambiguity

Assure robustness

Don’t be ignorant
Yeast Cell Cycle