Co-expression analysis of RNA-seq data

Etienne Delannoy & Marie-Laure Martin-Magniette & Andrea Rau

Plant Science Institute of Paris-Saclay (IPS2)
Applied Mathematics and Informatics Unit (MIA-Paris)
Genetique Animale et Biologie Integrative (GABI)
Outline

1. Co-expression analysis introduction

2. Unsupervised clustering
   - Centroid-based clustering: K-means, HCA
   - Model-based clustering
   - Mixture models for RNA-seq data

3. Conclusion / discussion
Aims for this talk

- What is the biological/statistical meaning of co-expression for RNA-seq?
- What methods exist for performing co-expression analysis?
- How to choose the number of clusters present in data?
- Advantages / disadvantages of different approaches: speed, stability, robustness, interpretability, model selection, ...
Design of a transcriptomics project

Biological question

Experimental design

Data acquisition

Data analysis:
Normalization, differential analysis, clustering, networks, ...

Validation

Co-expression analysis of RNA-seq data
Gene co-expression

Google image search: “Coexpression”
Gene co-expression is...

- The simultaneous expression of two or more genes\(^2\)
- Groups of co-transcribed genes\(^3\)
- Similarity of expression\(^4\) (correlation, topological overlap, mutual information, ...)
- Groups of genes that have similar expression patterns\(^5\) over a range of different experiments

\(^2\)https://en.wiktionary.org/wiki/coexpression
\(^3\)http://bioinfow.dep.usal.es/coexpression
\(^4\)http://coxpresdb.jp/overview.shtml
\(^5\)Yeung et al. (2001)
\(^6\)Eisen et al. (1998)
Gene co-expression is...

- The simultaneous expression of two or more genes
- Groups of co-transcribed genes
- Similarity of expression (correlation, topological overlap, mutual information, ...)
- Groups of genes that have similar expression patterns over a range of different experiments
- Related to shared regulatory inputs, functional pathways, and biological process(es)

\(^2\)https://en.wiktionary.org/wiki/coexpression
\(^3\)http://bioinfow.dep.usal.es/coexpression
\(^4\)http://coxpresdb.jp/overview.shtml
\(^5\)Yeung et al. (2001)
\(^6\)Eisen et al. (1998)
From co-expression to gene function prediction

- Transcriptomic data: main source of ’omic information available for living organisms
  - Microarrays (~1995 - )
  - High-throughput sequencing: RNA-seq (~2008 - )

Co-expression (clustering) analysis

- Study patterns of relative gene expression (profiles) across several conditions
- Co-expression is a tool to study genes without known or predicted function (orphan genes)
- Exploratory tool to identify expression trends from the data (≠ sample classification, identification of differential expression)
Let $y_{ij}$ be the raw count for gene $i$ in sample $j$, with library size $s_j$.

Profile for gene $i$:

$$p_{ij} = \frac{y_{ij}}{\sum_\ell y_{i\ell}}$$

Normalized profile for gene $i$:

$$p_{ij} = \frac{y_{ij}}{s_j}$$
Let $y_{ij}$ be the raw count for gene $i$ in sample $j$.

Profile for gene $i$:

$$p_{ij} = y_{ij} \sum_{\ell} y_{i\ell}$$

Normalized profile for gene $i$:

$$p_{ij} = \frac{y_{ij}}{s_j} \frac{\sum_{\ell} y_{i\ell}}{s_j}$$
Let $y_{ij}$ be the raw count for gene $i$ in sample $j$, with library size $s_j$

Profile for gene $i$: $p_{ij} = \frac{y_{ij}}{\sum_{\ell} y_{i\ell}}$
Normalized profile for gene $i$: $p_{ij} = \frac{y_{ij}/s_j}{\sum_{\ell} y_{i\ell}/s_j}$
Unsupervised clustering

Objective

Define **homogeneous** and **well-separated** groups of genes from transcriptomic data

What does it mean for a pair of genes to be close? Given this, how do we define groups?
Unsupervised clustering

Objective
Define **homogeneous** and **well-separated** groups of genes from transcriptomic data

What does it mean for a pair of genes to be close? Given this, how do we define groups?

Two broad classes of methods typically used:
1. Centroid-based clustering (K-means and hierarchical clustering)
2. Model-based clustering (mixture models)
Similarity measures

Similarity between genes is defined with a distance:

- **Euclidian distance** (L2 norm): \( d^2(y_i, y'_i) = \sum_{\ell=1}^{p} (y_{i\ell} - y'_{i\ell})^2 \)
  
  ⇒ Note: sensitive to scaling and differences in average expression level
Similarity measures

Similarity between genes is defined with a distance:

- **Euclidian distance (L2 norm):** 
  \[ d^2(y_i, y_i') = \sum_{\ell=1}^{p} (y_{i\ell} - y_{i'\ell})^2 \]
  Note: sensitive to scaling and differences in average expression level

- **Pearson correlation coefficient:** 
  \[ d_{pc}(y_i, y_i') = 1 - \rho_{i,i'} \]

- **Spearman rank correlation coefficient:** as above but replace \( y_{ij} \) with rank of gene \( g \) across all samples \( j \)

- **Absolute or squared correlation:** 
  \[ d_{ac}(y_i, y_i') = 1 - |\rho_{i,i'}| \] or 
  \[ d_{sc}(y_i, y_i') = 1 - \rho_{i,i'}^2 \]

- **Manhattan distance:** 
  \[ d_{\text{Manhattan}}(y_i, y_i') = \sum_{\ell=1}^{p} |y_{i\ell} - y_{i'\ell}| \]
Inertia measures

Homogeneity of a group is defined with an inertia criterion:

Let \( y_D \) be the centroid of the dataset and \( y_{C_k} \) the centroid of group \( C_k \)

\[
\text{Inertia} = \sum_{g=1}^{G} d^2(y_i, y_D)
\]

\[
= \sum_{k=1}^{K} \sum_{g \in C_k} d^2(y_i, y_{C_k}) + \sum_{k=1}^{K} n_k d^2(y_{C_k}, y_D)
\]

= within-group inertia + between-group inertia

ED & MLMM & AR

Co-expression analysis of RNA-seq data

INRA 11 / 42
In practice...

Objective: cluster $G$ genes into $K$ groups, maximizing the between-group inertia.

- Exhaustive search is impossible.
- Two algorithms are often used:
  1. K-means
  2. Hierarchical clustering

Co-expression analysis of RNA-seq data
K-means algorithm

Initialization $K$ centroids are chosen randomly or by the user.

Iterative algorithm

1. **Assignment** Each gene is assigned to a group according to its distance to the centroids.
2. **Calculation of the new centroids**

Stopping criterion: when the maximal number of iterations is achieved OR when groups are stable.

Properties

- Rapid and easy
- Results depend strongly on initialization
- Number of groups $K$ is fixed a priori
K-means illustration

Animation: http://shabal.in/visuals/kmeans/1.html
K-means algorithm: Choice of $K$?

- Elbow plot of within-sum of squares: examine the percentage of variance explained as a function of the number of clusters.

![Elbow plot example]

- Gap statistic: estimate change in within-cluster dispersion compared to that under expected reference null distribution.

- Silhouette statistic: measure of how closely data within a cluster is matched and how loosely it is matched to neighboring clusters.
Hierarchical clustering analysis (HCA)

Objective: Construct embedded partitions of \((G, G-1, \ldots, 1)\) groups, forming a tree-shaped data structure (dendrogram).

Algorithm:
- **Initialization**: \(G\) groups for \(G\) genes.
- **At each step:**
  - **Closest** genes are clustered.
  - Calculate **distance** between this new group and the remaining genes.
Distances between groups for HCA

Distances between groups

- **Single-linkage clustering:**

  \[
  D(C_k, C_{k'}) = \min_{y \in C_k} \min_{y' \in C_{k'}} d^2(y, y')
  \]

- **Complete-linkage clustering:**

  \[
  D(C_k, C_{k'}) = \max_{y \in C_k} \max_{y' \in C_{k'}} d^2(y, y')
  \]

- **Ward distance:**

  \[
  D(C_k, C_{k'}) = d^2(y_{C_k}, y_{C_{k'}}) \times \frac{n_k n_{k'}}{n_k + n_{k'}}
  \]

  where \( n_k \) is the number of genes in group \( C_k \)
Distances between groups for HCA

- Simple linkage
- Average linkage
- Complete linkage

Source: http://compbio.pbworks.com/w/page/16252903/Microarray%20Clustering%20Methods%20and%20Gene%20Ontology
HCA: additional details

Properties:
- HCA is stable since there is no initialization step
- $K$ is chosen according to the tree
- Results strongly depend on the chosen distances
- Branch lengths are proportional to the percentage of inertia loss
  $\Rightarrow$ a long branch indicates that the 2 groups are not homogeneous
Model-based clustering

- Probabilistic clustering models: data are assumed to come from distinct subpopulations, each modeled separately.
- Rigorous framework for parameter estimation and model selection.
- **Output**: each gene assigned a probability of cluster membership.

**What we observe**

**The model**

**The expected results**

\[ Z = ? \]

\[ Z : 1 = 
\[ , 2 = \]
\[ , 3 = \] \]
Key ingredients of a mixture model

- Let \( \mathbf{y} = (y_1, \ldots, y_n) \) denote the observations with \( y_i \in \mathbb{R}^Q \)
- We introduce a latent variable to indicate the group from which each observation arises:
  
  \[
  Z_i \sim \mathcal{M}(n; \pi_1, \ldots, \pi_K), \\
  P(Z_i = k) = \pi_k
  \]

- Assume that \( y_i \) are conditionally independent given \( Z_i \)
- Model the distribution of \( y_i | Z_i \) using a parametric distribution:
  
  \[
  (y_i | Z_i = k) \sim f(\cdot; \theta_k)
  \]
Questions around the mixtures

- **Model**: what distribution to use for each component?
  
  \( \Rightarrow \) depends on the observed data.

- **Inference**: how to estimate the parameters?
  
  \( \Rightarrow \) usually done with an EM-like algorithm (Dempster et al., 1977)

- **Model selection**: how to choose the number of components?
  
  A collection of mixtures with **a varying number of components** is usually considered
  
  A penalized criterion is used to select the best model from the collection
Clustering data into components

**Distributions:**

\[ g(x) = \pi_1 f_1(x) + \pi_2 f_2(x) + \pi_3 f_3(x) \]

**Conditional probabilities:**

\[ \tau_{ik} = \frac{\pi_k f_k(x_i)}{g(x_i)} \]

**Maximum a posteriori (MAP) rule:** Assign genes to the component with highest conditional probability \( \tau_{ik} \):

<table>
<thead>
<tr>
<th>( \tau_{ik} ) (%)</th>
<th>( k = 1 )</th>
<th>( k = 2 )</th>
<th>( k = 3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i = 1 )</td>
<td>65.8</td>
<td>34.2</td>
<td>0.0</td>
</tr>
<tr>
<td>( i = 2 )</td>
<td>0.7</td>
<td>47.8</td>
<td>51.5</td>
</tr>
<tr>
<td>( i = 3 )</td>
<td>0.0</td>
<td>0.0</td>
<td>100</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Model selection for mixture models

Asymptotic penalized criteria

- BIC aims to identify the best model $K$ wrt the **global fit** of the data distribution:

$$BIC(K) = - \log P(y|K, \hat{\theta}_K) + \frac{\nu_K}{2} \log(n)$$

where $\nu_K$ is the # of free parameters and $\hat{\theta}_K$ is the MLE of the model with $K$ clusters

- ICL aims to identify the best model $K$ wrt **cluster separation**:

$$ICL(K) = BIC(K) + \left( - \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \log \tau_{ik} \right)$$

$\Rightarrow$ Select $K$ that **minimizes** BIC or ICL (but be careful about their sign!)

---

[7] Asymptotic: approaching a given value as the number of observations $n \to \infty$
Model selection for mixture models: BIC vs ICL

![Cluster plot](image1)

- **BIC solution**: Shows clusters based on BIC criterion.

- **ICL solution**: Shows clusters based on ICL criterion.

![Criterion graph](image2)

- **Criterion value** plot comparing BIC and ICL for different numbers of clusters.

The plots illustrate the differences in cluster assignments and criterion values between BIC and ICL.
Model selection for mixture models
Non-asymptotic penalized criteria

Recent work has been done in a non-asymptotic context using the slope heuristics (Birgé & Massart, 2007):

$$SH(K) = \log P(y|K, \hat{\theta}_K) + \kappa \text{pen}_{shape}(K)$$

- In large dimensions, linear behavior of \( \frac{D}{n} \rightarrow -\gamma_n(\hat{S}_D) \)
- Estimation of slope to calibrate \( \hat{\kappa} \) in a data-driven manner (Data-Driven Slope Estimation = DDSE), *capushe* R package
Assume data $y$ come from $K$ distinct subpopulations, each modeled separately:

$$f(y|K, \psi_K) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(y_i; \theta_k)$$

- $\pi = (\pi_1, \ldots, \pi_K)'$ are the mixing proportions, where $\sum_{k=1}^{K} \pi_k = 1$
- $f_k$ are the densities of each of the components
Assume data $y$ come from $K$ distinct subpopulations, each modeled separately:

$$f(y|K, \psi_K) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(y_i; \theta_k)$$

- $\pi = (\pi_1, \ldots, \pi_K)'$ are the mixing proportions, where $\sum_{k=1}^{K} \pi_k = 1$
- $f_k$ are the densities of each of the components

For microarray data, we often assume $y_i|k \sim \text{MVN}(\mu_k, \Sigma_k)$

What about RNA-seq data?
Finite mixture models for RNA-seq data

\[ f(y|K, \Psi_K) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(y_i|\theta_k) \]

For RNA-seq data, we must choose the family & parameterization of \( f_k(\cdot) \):

1. Directly model read counts (HTSCluster):

\[ y_i|Z_i = k \sim \prod_{j=1}^{J} \text{Poisson}(y_{ij}|\mu_{ijk}) \]

2. Apply appropriately chosen data transformation (coseq):

\[ g(y_i)|Z_i = k \sim \text{MVN}(\mu_k, \Sigma_k) \]
Poisson mixture models for RNA-seq (Rau et al., 2015)

\[ y_i | Z_i = k \sim \prod_{j=1}^{J} \text{Poisson}(y_{ij} | \mu_{ijk}) \]

**Question:** How to parameterize the mean $\mu_{ijk}$ to obtain meaningful clusters of co-expressed genes?
Poisson mixture models for RNA-seq (Rau et al., 2015)

\[ y_i | Z_i = k \sim \prod_{j=1}^{J} \text{Poisson}(y_{ij} | \mu_{ijk}) \]

**Question:** How to parameterize the mean \( \mu_{ijk} \) to obtain meaningful clusters of co-expressed genes?

\[ \mu_{ijk} = w_i \lambda_{jk} s_j \]

- \( w_i \): overall expression level of observation \( i \) (\( y_i \)).
- \( \lambda_k = (\lambda_{jk}) \): clustering parameters that define the profiles of genes in cluster \( k \) (variation around \( w_i \)).
- \( s_j \): normalized library size for sample \( j \), where \( \sum_j s_j = 1 \)
Behavior of model selection in practice for RNA-seq

Contrast representation

The regression line is computed with 17 points

Validation points
Discussion of PMM for RNA-seq data

Advantages:

1. Directly models counts (no data transformation necessary)
2. Clusters interpreted in terms of profiles around mean expression
3. Implemented in HTSCluster package on CRAN (v1.0.8)
4. Promising results on real data...
Discussion of PMM for RNA-seq data

Advantages:
1. Directly models counts (no data transformation necessary)
2. Clusters interpreted in terms of profiles around mean expression
3. Implemented in HTSCluster package on CRAN (v1.0.8)
4. Promising results on real data...

Limitations:
1. Slope heuristics requires a very large collection of models to be fit
2. Restrictive assumption of conditional independence among samples
3. Cannot model per-cluster correlation structures
4. Poisson distribution requires assuming that mean = variance
Correlation structures in RNA-seq data

Example: data from Mach et al. (2014) on site-specific gene expression along the gastrointestinal tract of 4 healthy piglets
Gaussian mixture models for RNA-seq

Idea: Transform RNA-seq data, then apply Gaussian mixture models

Several data transformations have been proposed for RNA-seq to render the data approximately homoskedastic:

- $\log_2(y_{ij} + c)$
- Variance stabilizing transformation (DESeq)
- Moderated log counts per million (edgeR)
- Regularized log-transformation (DESeq2)

... but recall that we wish to cluster the normalized profiles

$$p_{ij} = \frac{y_{ij}/s_j}{\sum_\ell y_{i\ell}/s_j}$$
Remark: transformation needed for normalized profiles

Note that the normalized profiles are *compositional data*, i.e. the sum for each gene $p_i = 1$

This implies that the vector $p_i$ is linearly dependent $\Rightarrow$ imposes constraints on the covariance matrices $\Sigma_k$ that are problematic for the general GMM.

As such, we consider a transformation on the normalized profiles to break the sum constraint:

$$\tilde{p}_{ij} = g(p_{ij}) = \arcsin(\sqrt{p_{ij}})$$

And fit a GMM to the transformed normalized profiles:

$$f(\tilde{p} | K, \psi_K) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \phi(\tilde{p}_i | \theta_k, \Sigma_k)$$
Running the PMM or GMM for RNA-seq data with coseq

```r
> library(coseq)
>
> GMM <- coseq(counts, K=2:10, model="Normal",
transformation="arcsin")
>
> summary(GMM)
> plot(GMM)
>
> PMM <- coseq(counts, K=2:10, model="Poisson",
transformation="none")
>
> summary(PMM)
> plot(PMM)
```

---

ED& MLMM& AR

Co-expression analysis of RNA-seq data

INRA 35 / 42
Examining GMM results

Co-expression analysis of RNA-seq data
Examining GMM results

Co-expression analysis of RNA-seq data
Examining GMM results
Evaluation of clustering quality
Evaluation of clustering quality
Evaluation of clustering quality

![Graph showing clustering quality](attachment:graph.png)

- **Max conditional probability**
  - > 0.8
  - < 0.8

Cluster distribution:
- Cluster 9: 2000 observations
- Cluster 2: 1500 observations
- Cluster 7: 1000 observations
- Cluster 8: 500 observations
- Clusters 5, 10, and 3: 500 observations each
- Clusters 11, 12, 6, 4, and 1: 250 observations each
Conclusions: RNA-seq co-expression

Some practical questions to consider prior to co-expression analyses:

- **Should all genes be included?**
  Screening via differential analysis or a filtering step (based on mean expression or coefficient of variation)...
  ➞ Usually a good idea, genes that contribute noise will affect results!

- **What to do about replicates?**
  Average, or model each one independently?
  ➞ Note that the PMM makes use of experimental condition labels, but the GMM does not...
Clustering results can be evaluated based on internal criteria (e.g., statistical properties of clusters) or external criteria (e.g., functional annotations).

Preprocessing details (normalization, filtering, dealing with missing values) can affect clustering outcome.

Methods that give different results depending on the initialization should be rerun multiple times to check for stability.

Most clustering methods will find clusters even when no actual structure is present ⇒ good idea to compare to results with randomized data!

---

D’haeseller, 2005
Difficult to compare several clustering algorithms on a given dataset (and difficult to discern under which circumstances a particular method should be preferred)

- **Adjusted Rand index**: measure of similarity between two data clusterings, adjusted for the chance grouping of elements
  - $\text{ARI}$ has expected value of 0 in the case of a random partition, and is bounded above by 1 in the case of perfect agreement
A note about validating clustering approaches on real data

- Difficult to compare several clustering algorithms on a given dataset (and difficult to discern under which circumstances a particular method should be preferred)
  - **Adjusted Rand index**: measure of similarity between two data clusterings, adjusted for the chance grouping of elements
    - ARI has expected value of 0 in the case of a random partition, and is bounded above by 1 in the case of perfect agreement

- Difficult to evaluate how well a given clustering algorithm performs on transcriptomic data

- No one-size-fits-all solution to clustering, and no consensus of what a “good” clustering looks like ⇒ use more than one clustering algorithm!
There is no single best criterion for obtaining a partition because no precise and workable definition of *cluster* exists. Clusters can be of any arbitrary shapes and sizes in a multidimensional pattern space. Each clustering criterion imposes a certain structure on the data, and if the data happen to conform to the requirements of a particular criterion, the true clusters are recovered.

---

*Jain & Dubes, 1988*
Acknowledgements & References

MixStatSeq ANR-JCJC grant

Thanks to Gilles Celeux (Inria Saclay - Île-de-France), Cathy Maugis-Rabusseau (INSA / IMT Toulouse), Etienne Delannoy, Marie-Laure Martin-Magniette (SPS), and Panos Papastamoulis (University of Manchester)
