



Similarity measures in clustering time series data

Paula Silvonen
paula.silvonen@vtt.fi

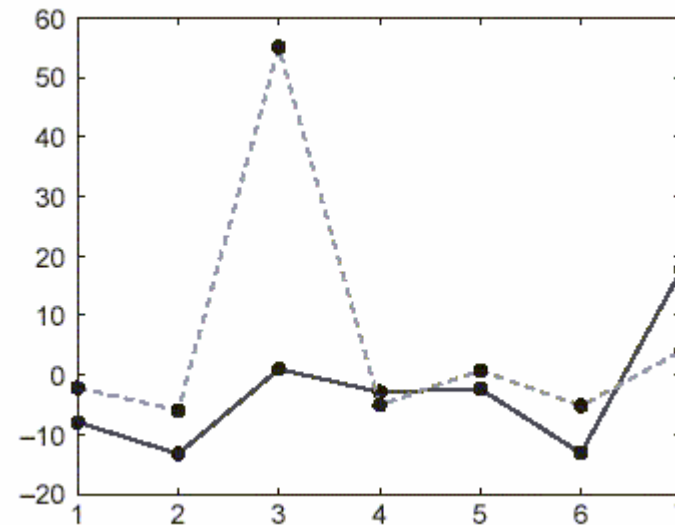
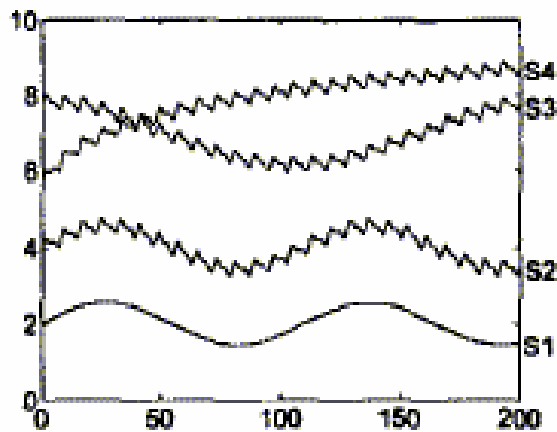


Introduction

- Clustering:
 - determine the similarity or distance between profiles
 - group the expression profiles according to their similarities
- Time series data: measurements at various time points in a sequence, a curve of a set of measurements as a function of time
- Standard similarity measures: Euclidian distance, Pearson correlation, ...
- A relationship among genes at the biological level often presents itself by **locally** similar and potentially **time-shifted** patterns in their expression profiles
- Similarity degrees between genes by comparing complete expression profiles may not be informative
- Similarity degrees between genes by comparing values at the same time point will not reveal "master-slave" relationships

Introduction

- Filtering the data to contain only significant changes may remove transcriptional regulators
- What do we wish to analyse in the data?



Simple solution?

- Enumerate all possible alignments of expression profiles
- Compute similarity of two profiles by sliding two windows over them
- For long expression profiles computationally too expensive
 - => Some kind of approximation or data transformation is needed

Using partial information (Jin et al.)

- A representing model for the partial information:
<F, T, D> where F is a decomposition method, T a representation method, D a distance measure
- For example, orthonormal discrete transform and Euclidian distance
- Retrieve the partial information and store it in a set of components
- Represent them in a compact form
- Add weight A_k to the components on the basis of interest
- Measure the distance of partial information

Using partial information

- Example:
 - F: Decompose the time series to two components: local fluctuation, global movement
 - T: Map fluctuation components to α and 0
 - If we are interested in the local movement, let $K=(1)$ and $A_k=(1,0)$
 - D: Euclidian distance

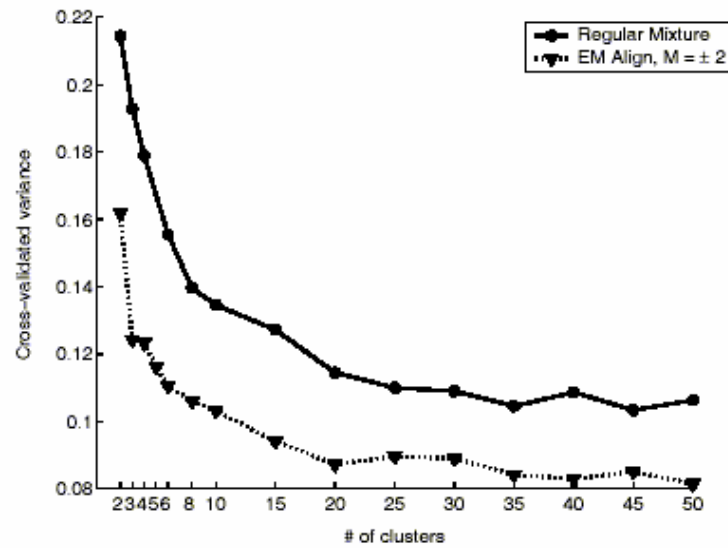
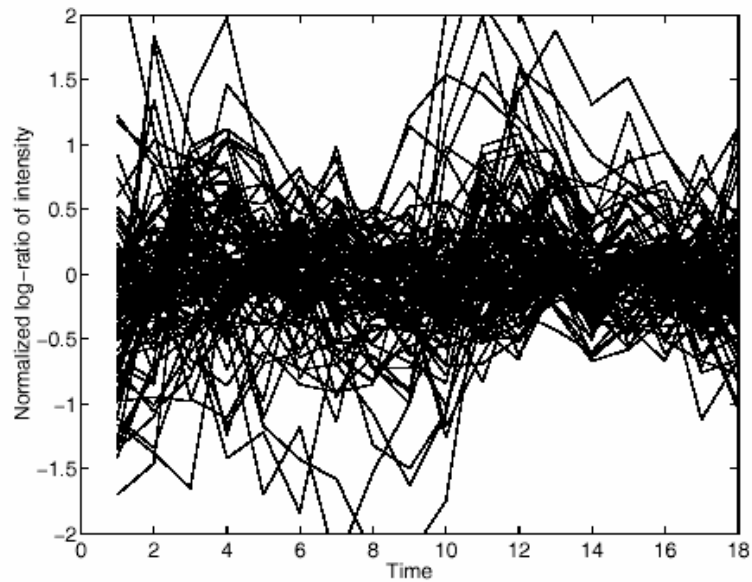
Time curve clustering with mixture models (Chudova et al.)

- Simultaneous clustering and alignment of sets of curves
- Discrete-valued shifts along the time (or independent variable) axis and real-valued additive offsets in each of the measurement (dependent variable) axes
- Curves are assumed to be generated from a particular class of generative models
- A curve is represented as a point in T-dimensional space, T is the length of the longest curve + the maximum allowed time shift
- Distribution modelled as a multivariate Gaussian with a diagonal covariance matrix
- Latent variables Z_i (cluster membership) and φ_i (amount of shifting on the time grid)

Time curve clustering

- Parameters are estimated with EM (Expectation maximization)
- **E-Step**: evaluate the distribution of latent variables Z_i and φ_i given current parameter estimates
- **M-Step**: adjust free parameters of the model to maximize the expected log-likelihood of the data with respect to the distribution of latent variables
- Hierarchical Bayesian model for smoothing the mean curves, conjugate Gamma priors for the diagonal covariance terms in the Gaussian mixture components and Dirichlet priors for the mixture component probabilities and time shift probabilities within the clusters

Time curve clustering



q -clusters (Ji and Tan)

- Identify localized time-lagged co-regulations between genes and/or gene clusters
- Genes in the same cluster have a similar expression pattern over q consecutive conditions
- Pattern describes a changing tendency, which reflects how the expression value changes from a condition to the next for the q conditions
- Types of co-regulation: (1) genes with the same starting time point (2) genes activate genes with later starting time point
- Co-regulations/inhibitions across clusters: (1) genes inhibit genes with later starting time point and complementary expression pattern (2) genes activate genes with later starting time point and similar expression pattern

q -clusters

- Transform the expression matrix into a "slope" matrix to reflect the genes' changing tendency
- Each entry shows the directional change from an expression value to the next (1, -1, 0)
- Genes with the same expression pattern in some q consecutive time points are clustered together (exact match)
- Form biclusters with a mean-squared residue metric
- Identify possible activation co-regulations
- Identify possible inhibition regulations
- Add approximate matching to find more possible co-regulators
- Simple and efficient, but is enough information preserved?

CLARITY (Balasubramaniyan et al.)

- A shape-based similarity measure based on the Spearman rank correlation
- A compromise between numerical measures like Pearson correlation and simple qualitative measures?
- Enumerate all possible alignments of expression profiles
- Find an initial "hit" in the form of a short optimal alignment, if not unique, handle all candidates
- Extend the hits in both directions, neighbourhood size= d , find best match
- Iterate until the optimal alignment does not change

CLARITY

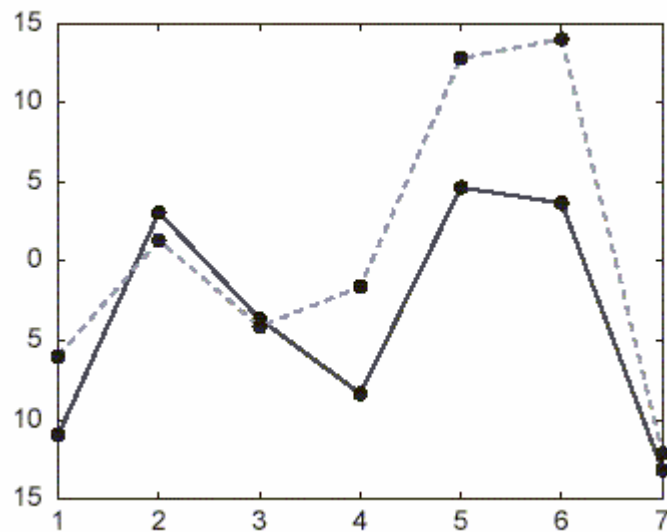
- Spearman rank correlation retains more information than qualitative measure that compares the changing tendencies

$$\text{SRC}(X, Y) = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n (r_X(x_i) - r_Y(y_i))^2,$$

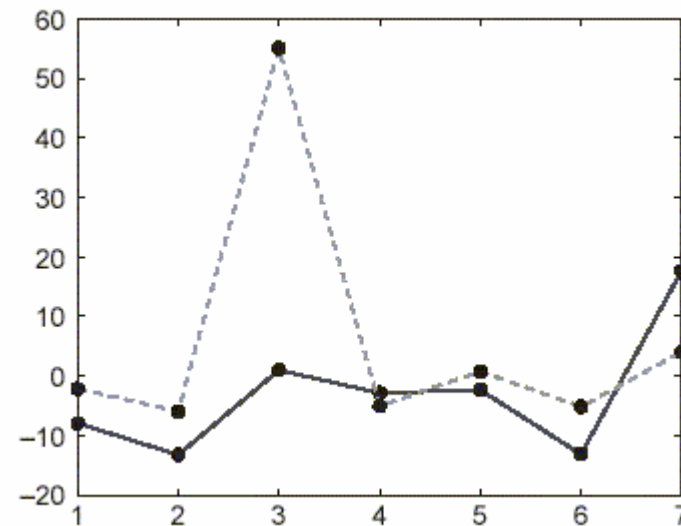
where $r_X(x_i)$ is the rank of x_i in the profile $(x_1 \dots x_n)$: $r_X(x_i) = k \iff |\{j \mid x_j < x_i\}| = k - 1$, $k = \text{length of the alignment}$

CLARITY

- Takes into account the shape of the profiles



Qualitative measure 0.3, SRC 0.93



Pearson 0.3, SRC 0.93

Conclusions

- The selection of a similarity measure is of crucial importance in clustering
- Time series data requires special handling if we wish to find (possibly time-lagged) similar subsequences
- Local similarities
- Time shifting
- Different models for representing the data

References

- Rajarajeswari Balasubramaniyan, Eyke Hüllermeier, Nils Weskamp and Jörg Kämper. Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics* Vol. 21 no. 7 2005, pages 1069–1077.
- Liping Ji and Kian-Lee Tan. Identifying time-lagged gene clusters using gene expression data. *Bioinformatics* Vol. 21 no. 4 2005, pages 509–516.
- Darya Chudova, Scott Gaffney, Eric Mjolsness, Padhraic Smyth. Translation-Invariant Mixture Models for Curve Clustering. In *Proceedings of SIGKDD 03 August 24-27, 2003 Washington D.C., USA*, pages 79-88.
- Xiaoming Jin, Yuchang Lu, Chunyi Shi. Similarity Measure Based on Partial Information of Time Series. In *Proceedings of SIGKDD 02, July 23-26, 2002, Edmonton, Alberta, Canada*, pages 544-549.