Extending an Algorithm for Clustering Gene Expression Time Series

Mikko Korpela and Jaakko Hollmén
Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland

Introduction

Many interesting phenomena occur as dynamic responses to the experimental conditions. Short time series with less than 10 time points are typical in gene expression studies.

Clustering short time series data

Short time series can be represented as vectorial data, where different time points are represented by the elements of the vector. Clustering can reveal typical dynamic gene expression profiles in the data. We have used [1] an algorithm for short time series clustering presented in [2] and extended the algorithm. Fig. 1 illustrates the data representation and the clustering solution using the approach. The stages of the clustering algorithm are outlined in the *blue box* on the right.

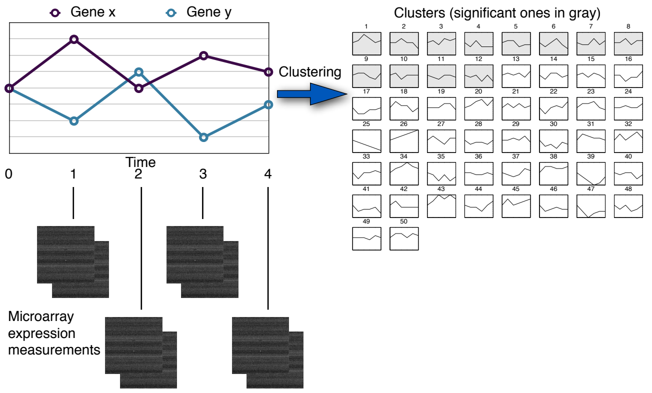
Gene x
Gene y
Time
0
1
2
3
4
Clustering
Clusters (significant ones in gray)
Microarray expression measurements

Fig. 1: Time series is formed from gene expression measurements (left), the clustering solution (right)

Improvements to the algorithm

Our corrections and extensions to the clustering algorithm (stages 1 and 2) are listed in the *green box* on the right. The main change is the revised profile selection algorithm (Algorithm 3) that addresses the problem shown in Fig. 2. Table 1 is an example of the improvements attainable with randomization as used in Algorithm 3.

Clustering short time series [2]

1. Enumerate all cluster prototypes (expression profiles) of a certain (discrete) kind
2. From that set, select a number of prototypes
3. Assign each gene to a cluster
4. Assess statistical significance of clusters
5. (Optionally) divide clusters into groups

New!

Improvements

- Remove zero profile to solve issue with correlation
- Remove redundant profiles before profile selection
- Fix profile selection with randomized algorithm (Fig. 2)

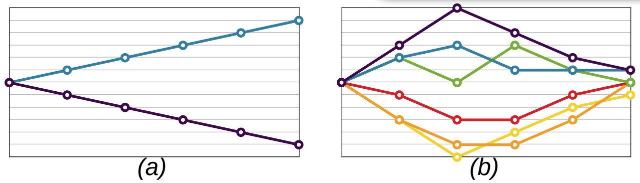

Fig. 2: Ambiguity in greedy profile selection. (a) The first two profiles (b) Some equally good choices when choosing the third profile (randomization can be used)

Type of algorithm		Parameters				
	n	5	6	5	6	3
	c	2	2	3	3	6
	m	50	50	50	50	16
Deterministic		0.1548	0.2572	0.1784	0.2843	0.0695
Randomized		0.1708	0.2929	0.1982	0.2960	0.0695

Table 1: Minimum distance between selected profiles (larger is better). Randomized algorithm (Algorithm 3) provides improved results. Results were achieved with 100 (in the "n=6, c=3" case) or 1000 repeats (parameter in Algorithm 3). Parameters: n = length of profiles, c = maximum change between time points (affects total number of profiles), m = number of profiles selected.

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}, \quad a_i = x_i - \overline{x},\ b_i = y_i - \overline{y}$$

is not defined when $\boldsymbol{x}$ or $\boldsymbol{y}$ is a constant (zero) profile

Algorithm 2 RemoveRedundant A simple algorithm for removing redundant model profiles

```
RemoveRedundant($P, c, n$)
$R \leftarrow \{\}$
let $Primes$ be the set of all prime numbers up to and including $c$
while $|P| > 0$
    do let $p$ be any profile in $P$
       $P \leftarrow P \setminus \{p\}$
       $nonredundant \leftarrow$ TRUE
       let $p_i$, $i \in 1, \ldots, n$, be the values at each time point of $p$
       for each $prime$ in $Primes$
           do if each $p_i$ is divisible by $prime$
                 then $nonredundant \leftarrow$ FALSE
                      break
       if $nonredundant$
          then $R \leftarrow R \cup \{p\}$
return $R$
```

Algorithm 3 SelectVectorsMaxMinDistRandom A randomized greedy algorithm for choosing m distinct profiles

```
SelectVectorsMaxMinDistRandom($d, P, m, repeats$)
$dist_{best} \leftarrow -\infty$
for $i \leftarrow 1$ to $repeats$
    do $R_t \leftarrow$ SelectHelper($d, P, m$)
       $dist_{temp} \leftarrow \mathbf{min}_{(p_1,p_2) \in R_t \times R_t} d(p_1, p_2)$
       if $dist_{temp} > dist_{best}$
          then $dist_{best} \leftarrow dist_{temp}$
               $R \leftarrow R_t$
return $R$

SelectHelper($d, P, m$)
let $p_1 \in P$ be the profile that always goes down one unit between time points
$R \leftarrow \{p_1\}$
$L \leftarrow P \setminus \{p_1\}$
for $i \leftarrow 2$ to $m$
    do let $p \in L$ randomly be one of the profiles that maximize $\mathbf{min}_{p_1 \in R} d(p, p_1)$
       $R \leftarrow R \cup \{p\}$
       $L \leftarrow L \setminus \{p\}$
return $R$
```

References

1. Korpela, M.: Analysis of changes in gene expression time series data. Master's thesis, Helsinki University of Technology, Finland (February 2006)
2. Ernst, J., Nau, G.J., Bar-Joseph, Z.: Clustering short time series gene expression data. Bioinformatics 21(Suppl. 1) (2005) i159–168

E-mail: mvkorpel@cis.hut.fi , Jaakko.Hollmen@hut.fi