Introduction to Computational Biology
Lecture # 29: Clustering gene expression data

Keren Yizhak
23/06/2008

1 A Brief Review

Last lesson we talked about a method which performs clustering over genes and over arrays (experiments). This method creates homogenous clusters.

In this lesson we will talk about another method which looks for clusters of subsets of genes and experiments.

2 Bi-Clustering

This method tries to find a subset of genes and subset of experiments which share a similar behavior under certain conditions. We are actually looking for a square (as can seen in Figure 1) - only a subset of genes and arrays which has interesting connections between them.

How can we find such a subset?

Figure 1: A subset of genes and arrays.
1. Define a score, try all subsets of genes and experiments and take those who pass this score.

2. Assuming we know the subset of genes, we can iterate over the arrays and check which one shows an interesting behavior over this subset. If the subset of the arrays is known we can do the same for the genes.

Let’s define an iterative procedure which solve this problem.

2.1 The Signature Algorithm

- guess \( G^0 \)
- loop: \( t = 1 \) until convergence
  - \( A^t = F(X,G^{t-1}) \)
  - \( G^t = F(X,A^t) \)
- return \( A^t, G^t \)

\( X \) is the gene expression matrix, \( G^0 \) is the initial subset of genes and \( F \) is the following operator:

\[
F(X,G) = \{a : \# \{ |g \in G : |X_{g,a}| > (\mu_a + \sigma_a^2 * \alpha)\} > \beta * |G| \}
\]

\( F \) checks if a subset of genes \( G \) which every \( g \) in it is greater than \( (\mu_a + \sigma_a^2 * \alpha) \) is greater than a certain threshold: \( \beta * |G| \).

Now, there are two possibilities:

1. One of the iterations returns an empty subset. It might happen if we chose a problematic initial subset of genes.

2. \( A^t = A^{t-1} \rightarrow convergence \).

In order to deal with the first possibility we can choose number of initial subsets and run the algorithm. We will therefore get number of answers. This is possible because there isn’t necessarily just one correct answer. We can now compare those answers with the values of the matrix and check if we could get those clusters by chance. The answer will probably be no.

2.2 Graph Theoretic Method

We can solve the same problem with the same iterative idea but as a problem in graph theory.

Let’s create a bipartite graph. It will have two kinds of nodes, at one side it will have gene nodes and on the other side it will have array nodes. We will define an edge between a gene and an array if the expression of this gene in that array is interesting. We can see such a graph in Figure 2.
Now, as we defined the problem, we would like to find a heavy subgraph (looking for a clique will be too stringent). This means that we expect that every gene will have a high expression (or any other unique quality) in almost all arrays and every array will affect almost all genes. This problem is NP-hard and therefore we are not promised to get a solution.

In order to solve this problem we can use the same simple idea from the previous section: we’ll define an initial subset of genes, check to which array nodes it connects and so on iteratively.

2.3 Summary

In the two methods we described above we might get an overlap between clusters because there is a possibility that a subset of genes participates in a number of conditions and therefore should appear in a number of clusters.

We learned on global two-sided clustering and on Bi-Clustering, each has its own advantage: the global clustering is based on a lot of sets of genes and thus helps divide the arrays into a lot of conditions. On the other hand, the Bi-Clustering method allows overlapping and therefore is more flexible.

We would like to find something between those two methods, meaning, to get a small number of subsets from the genes and from the array groups. We can do that by changing the parameters which control the density and by that get different levels of accuracy. For tight parameters (such as looking for very unique expressed genes) we will get smaller clusters and for less tight parameters the clusters will get larger. This will allow us to slowly increase our resolution level.
3 Biological Networks

Up until now we looked for interesting phenomena in our data. Now, we would like to understand the mechanisms which cause these phenomena.

One of the mechanisms which we can learn from gene expression experiments is transcriptional regulation. We can identify that this kind of regulation is involved if we see a change in the gene expression from 0 to a high level.

The question we ask is what in the transcription regulation mechanism explains the co-expression of the genes. We can answer it in two ways:

1. From Expression to Regulation. We will take the gene expression data, cluster genes based on expression and then look for enrichment motifs in promoters of each cluster. Let's notice that in this stage we still don't know how these candidates work.

2. From Regulation to Expression. We will find a motif and check if the genes which it appears on create a cluster. In order to look for a motif we will take all subsets of sequences with some consensus size (for example, for sequences with 6 nt we will have $4^6$ possibilities). We will look for each one of the possible motifs in all the promoters and then check if there is a subset of genes that include a certain motif and express in a similar way. If such a subset is found we can try to expand the motif or maybe change it a little and then repeat the search.

The main signals will be found in both ways.

By EM we can join these two methods.

We need to find a function which gets a promoter sequence and maps it to an expression vector suitable for the 150 conditions -

$$f\text{(promotor)} \rightarrow \text{(expVector)}$$

But, we are actually looking for binding sites in a promoter and there are many possibilities for that. There might be a specific transcription site $X$, there might be an overlap between two transcription sites $X$ and $Y$, or a transcription site $Z$ which when changed causes a loss of the regulation.
The simplest function however is the one which says if the motif occur in a promotor or not.
Intuitively, we can solve this problem like that:
We have Expression, Motifs and Bindings (a promoter has a certain motif or not) as described in Figure 4.

We will find a set of motifs as we did before. Then we define a score which decide whether a promotor contains this motif or not. Now, with those two in hand we can check the Expression and look for interesting subsets. We can also go now in the other direction from Expression, fixing our function according to previous results and finding the Binding and the Motifs.
Let’s define random variables:

- $X$ - Expression
- $S$ - Promotor sequence
- $B$ - Binding

$X$ and $S$ are known and $B$ is a latent variable.

We also define two parameters:

- $\theta_f$ - parameter of $f$
- $\theta_m$ - motif parameter

We can describe all that in a model:

$$P(X, S, B) = P(X|B, \theta_f) \ast P(S|B, \theta_m) \ast P(B)$$

We can now run EM.

Let’s notice that in this procedure we only dealt with the gene axis and not with the array axis.