Gene expression, analysis of differential expression, co-expression

1 Background

Genes are the hereditary units of biological organisms. They are encoded into DNA residing in the chromosomes that are in the nucleus in eukaryotes (= cell contains a nucleus and other subdivisions), and more freely floating within bacteria. Each chromosome essentially contains a long DNA chain, consising of a sequence formed of four amino acids, A, C, G, and T. A gene is a subsequence within this long sequence. The so-called coding region of a gene is formed of one or more consecutive sequences (exons) possibly intervened with sequences not belonging to the gene (introns). Before the start of the coding region there is a promoter region which plays a role in starting and regulating the activity of the gene, and there may be numerous enhancer locations scattered around in the vicinity of the gene. When suitable proteins are bound to the enhancers, the activity of the gene is enhanced.

When genes are active a blueprint of their coding sequences is made and transferred out of the nucleus to *ribosomes* where the protein is to be manufactured. The blueprint is made in the form of a messenger molecule mRNA (for messenger RNA) which is essentially a copy of the coding sequence. In the ribosomes the mRNA sequence is read three letters, a *codon*, at a time. Each triplet corresponds to one out of 20 amino acids, which are building blocks of proteins. The amino acids are put to a sequence in the order determined by the mRNA sequence, and then the sequence *folds* into its natural low-energy configuration, a *protein.* Finally, several of these chains may form a protein complex.

A fundamental property of DNA and RNA molecules is that they bind together if their sequences are suitable. In DNA, letter A binds to T and letter C to G. In RNA the letters are slightly different but the same principle applies. For instance, AGTC binds with TCAG. As is well known, the chromosome is formed of a double-stranded DNA chain, meaning that there are two corresponding chains bound together. Such tendency for binding is very useful in constructing measurement devices: we can construct a *probe*, a short sequence specific to a certain gene or region of chromosome and expose it to mRNA or DNA samples (processed suitably). If there are matching sequences around, the probe is likely to bind to them.

The amount of mRNA molecules being produced of a certain gene depends on various molecules being active and available around the chromosomal region of the gene. Most notably, there needs to be a transcription initiator complex bound to the promoter region of the gene. A molecule, RNA polymerase, belonging to this complex then does the actual encoding of the DNA into mRNA. The rest of the proteins are called transcription factors (TF); they need to be bound to the DNA sequence in suitable locations. There exist several types of TFs, and they bind to different kinds of DNA sequences. The activity of the gene depends on the configuration of TFs that are bound to the potential binding sites close to the gene.

All these proteins are products of some other genes being first transcribed and then translated. All genes together form a dynamical system, a *regulatory network*, driven by its own dynamics and signals coming from outside of the cell.

Activity of a gene additionally depends on the configuration of the chromosome; a region can be silenced by adding certain compounds to the DNA molecule (process called methylation). Small RNA molecules transcribed elsewhere in the genome may bind to the mRNA molecule before it reaches a ribosome, stopping it from being translated into a protein. And finally, the resulting aminoacid chain may be modified by additional reactions to become more or less active, or active in a different way. All these and other mechanisms add to the complexity of the regulatory machinery of the cell, and the list is not comprehensive.

The (relative) amount of mRNA molecules stemming from a certain gene can be used as a measure of the activity of the gene, or *gene expression.* The amount can be measured by designing a probe sequence, that is, a sequence that is complementary to mRNA from the gene, and specific to that gene. When mRNA solution is then washed over such probes, the mRNA specific to the probe becomes bound or *hybridized* to the probes[1]. A breaktrough in functional genomics studies has been the invention of gene expression microarrays, that is, arrays containing probes for a large number of genes. Such arrays can be used to simultaneously measure expression of even tens of thousands of genes.

There are two main varieties of gene expression microarrays. Spotted cDNA arrays are glass slides on which a small amounts of solution containing probes are printed by physically dipping printing tips into wells containing the solution, and then touching the slide. Each resulting drop contains probes for a certain gene. The probes are sequences complementary to the whole mRNA of the gene or parts of it.

The amount of mRNA bound to a specific probe type is read from the slide with the help of a laser and fluorescent dyes. The mRNAs of a sample have been labeled by attaching fluorescent dyes to them, and when a laser beam with a wavelength matching the dye is aimed at the slide, the emitted intensity at each location reveals the number of mRNA molecules bound there.

To control for some of the numerous sources of noise, two samples, case and control, are typically hybridized simultaneously on the slide. Each sample is labeled with a different dye, and the (log) ratio of the intensities then reveals the relative amount of mRNA from the two samples.

The second type is oligonucleotide microarrays, where the probes (oligonucleotides = short, typically synthesized DNA or RNA sequences) are manufactured using litographic techniques directly according to a desired target. The probes are shorter but can be designed to be specific, and the step of having to (biologically) produce a collection (library) of cDNA needed in the cDNA arrays can be skipped. The oligonucleotides can either be directly grown in desired quantities on an array, or spotted on slides in the same way as the cDNA.

There are several significant sources of noise in the measurement process. Biological noise, stemming from sampling the cells and the treatments the cells undergo, can only be avoided through careful biological experimental design. Sample preparation is another; a large number of cells need to be pooled to get enough mRNA, which means that inhomogeneities in sampling the cells or synchronizing them will introduce additive noise. Laboratory practices may vary; for instance non-uniformities in washing of the liquid over the slide may cause noise. The binding affinity, the binding strength of the probes may vary, both over different slides in spotted cDNA microarrays and over different genes in all arrays. The dyes may affect the binding likelihood, and different dyes to different extents.

Some of the noise can be reduced though standardized laboratory practices and processes. For the rest we need statistical noise models and/or normalizing preprocessing. If measurements have been replicated (which naturally is a good idea), we can assume that variation between replicates is noise, and making further assumptions such as that noise is independent across genes and samples, maybe normally distributed (which may not be realistic, though), and even that the signal may be shared by groups of genes, the underlying signal can be estimated together with its confidence intervals.

Alternatively, if known amounts of certain mRNA samples have been injected into the samples, a procedure called “spike-in”, deviations from the known value can be assumed to be due to experiment-specific biases and scaling which can be estimated. Assuming the biases and scaling are not specific to the spiked-in controls, they can be removed from the other genes as well.

If there are no or very few replications, more drastic assumptions need to be made to normalize the expression measurements. For instance, the sample mean or some other statistcs of the distribution of expression can be assumed unreliable and removed to make different chips more commensurable.

[1] technically, everything is first converted from RNA to DNA, called complementary DNA or cDNA

Alternatively, the mean and variance of each gene may be assumed non-informative or reliable, and removed removed by applying the z-tranformation.

In this course we will not go into more details of preprocessing or normalization, but we will return to the problem of controlling for uninteresting variation when talking about differential expression.

For now, assume that the gene expression data matrix $\mathbf{X}$ is, after preprocessing, a reasonably good representation of the activity of genes. The matrix consists of gene expression profiles $\mathbf{x}_i$ for gene i, such that x_{ij} is the expression level of gene i in treatment (microarray j).

2 Clustering and co-expression

Gene expression data has been used to infer functional relationships between the genes. If two genes are expressed in the same way, that is *co-expressed*, it is more reasonable to believe that their function is related than if they are expressed very differently or even independently. Co-expression may be due to co-regulation, meaning that their activity is determined by (partly) the same molecular mechanisms. Knowing the regulatory mechanisms would naturally be great but based on co-expression only we unfortunately cannot tell whether the reason is co-regulation or some outside factor that affects both of the co-expressed genes. The hypothesis of related function or even of co-regulation naturally becomes the stronger the wider the range of conditions in which the genes are co-expressed.

A simple way of assessing co-expression of genes i and j is to compute the correlation $c_{ij} = \sum_k (x_{ik} - \mu_i)(x_{jk} - \mu_j)/\sigma_i/\sigma_j$, where the means and variances are estimated over the set of conditions k. If the correlation is higher than expected by chance, there is some relationship between the genes. Given a reasonable number of conditions, significance could be estimated by a permutation test for instance. Assess whether the observed c_{ij} is exceptionally high among the set of data obtained by randomly permuting the indices k for one of the genes.

Correlation is of course only a crude measure of dependency; other estimates of mutual information have been used as well.

The pairwise correlations between all gene pairs form a network, a graph where all gene pairs are connected with links having the weight of their correlation. The graph only considers two-way interactions and still for N genes there are N^2 potential links. It would be nice to summarize.

2.1 Hierarchical clustering

Groups of more than one co-expressing gene could be found with the following algorithm: Take the two genes with the highest correlation in their expression, say i and j, and decide that they are co-expressed at the level c_{ij}. This is the first group or *cluster* consisting of genes i and j. Then continue by again adding together the two genes having the highest correlation, or a gene and a cluster if their correlation is the highest; defining the correlation between a gene and a cluster to be the minimum of all pairwise correlations ascertains that if the process is stopped at a certain level, all genes within any cluster have co-expression higher than the stopping level. Then this algorithm is continued, at each step adding together two genes, a gene and a cluster, or two clusters, until only one cluster remains.

The process can be visualized as a tree, a *dendrogram* [see the net for figures...], where the progressive branches correspond to the clusters and the leaves in the bottom are the genes. The tree can be cut at a desired level to get a fixed set of clusters.

This algorithm is called *hierarchical clustering*. In general, clustering is the task of finding groups of mutually similar data items, such that the different groups are relatively dissimilar. The task requires choosing a measure for similarity, and possibly also some other parameters that govern what kinds of clusters are to be sought. We used correlation as the distance (actually similarity) measure, choosing to group together genes with correlated expression. Additionally we chose that clusters need to be relatively compact, by requiring all pairs of genes within a cluster to have a high correlation, instead of defining the correlation between a cluster and a gene to be the highest correlation among the cluster (resulting in so-called single-link clustering, instead of the complete-link clustering here).

Some of the main pluses of hierarchical clustering, among the numerous clustering algorithms, are that both the algorithm and the resulting dendrogram visualization are intuitively understandable. Some minuses are that the result may be sensitive to noise, and that being defined as an algorithm instead of a well-defined objective, deep down its functioning may be hard to predict or control. But it functions well in its task of grouping together similar data items and visualizing the groups.

2.2 K-means and model-based clustering

An alternative, equally traditional clustering method is to choose the number of clusters and the distance measure, and then search for a set of clusters that are maximally homogeneous in terms of that distance measure. In K-means the clusters are defined by a set of *cluster prototypes* or centroids, such that all samples belong to the cluster having the closest centroid, in terms of the chosen distance measure. Clusters are homogeneous or contain similar samples, if the distance from each sample to its cluster centroid is minimal. Hence, the cost function to be minimized is

$$\sum_i \min_{z(i)} ||\mathbf{x}_i - \boldsymbol{\mu}_z||^2 = \sum_i ||\mathbf{x}_i - \boldsymbol{\mu}_{z(i)}||^2 \tag{1}$$

where $z(i)$ is the index of the closest centroid.

The cost function can be minimized by alternating two optimization steps: First compute $z(i)$ for each i, and then minimize (1) for the fixed set of z's, yielding, by taking the root of the gradient,

$$\boldsymbol{\mu}_k = \frac{\sum_{z(i)=k} x_i}{\sum_{z(i)=k} 1} \; .$$

K-means algorithm is (relatively) fast and intuitively readily understandable, and it is easy to interpret the results since the cost function has been designed to directly capture the intuition of clusters as mutually similar sets of samples. What is less pleasing is that the number of clusters needs to be chosen beforehand, and only more or less heuristic methods are available for aiding in that choice. A feature of K-means is that it favors, by its nature of minimizing the distance from cluster centroids, clusters that are spherical according to the given distance measure. This may or may not be desirable.

A very closely related alternative is to assume that the data points have been generated by a set of K generators, each generating data according to a specific distribution, typically Gaussian. The parameters of the distributions are unknown, and have to be estimated from the data. Assuming normally distributed generators with the same variance in each treatment, the kth generator produces data according to the distribution $p_k(\mathbf{x}|\boldsymbol{\mu}_k, \sigma_k) \sim N(\mathbf{x}|\boldsymbol{\mu}_k, \sigma_k \mathbf{I}) = \prod_j N(x_j|\mu_{kj}, \sigma_k)$. The whole mixture model produces the distribution

$$p(\mathbf{x}|\theta) = \sum_k p_k p(\mathbf{x}|\boldsymbol{\mu}_k, \sigma_k) \; ,$$

where θ is a shorthand notation meaning all parameters of the model.

The model can be interpreted to have two sets of unknowns, first the parameters $\boldsymbol{\mu}$ and σ, and secondly the knowledge of which generator generated each data point. The latter can be interpreted as *latent variables*, a discrete variable for each data point, which moreover can be considered as missing data. The EM (Expectation Maximization) algorithm is a widely used tool for fitting models where data is missing. It *maximizes the likelihood* of data D, $p(D|\theta) = E_{\mathbf{z}}[p(D, \mathbf{z}|\theta)]$, where $D = \{\mathbf{x}_i\}$ and the $\mathbf{z}$ contains all the z_i.

In the first, E-step, the distribution of latent variables given θ (the old, fixed value) is computed. Here

$$P(z_i = k|\mathbf{x}_i, \theta) = \frac{P(z_i = k, \mathbf{x}_i|\theta)}{p(\mathbf{x}_i|\theta)} = \frac{p_k P(\mathbf{x}_i|\mu_k, \sigma_k)}{P(\mathbf{x}_i|\theta)} \ .$$

In the second, M-step, the (probabilistic) assigments of samples to clusters are fixed to the values $p(z_i = k|\mathbf{x}_i, \theta)$ obtained in the E-step, and the (log) likelihood is maximized with respect to the parameters p_k, $\boldsymbol{\mu}_k$ and σ_k. Now the cluster assignments do not depend on the parameters any more, and the generating distribution of each cluster can be estimated separately. For normal distributions, the formula for the mean parameters,

$$\mu_k = \frac{\sum_i p(z_i = k|\mathbf{x}_i, \theta) x_i}{\sum_i p(z_i = k|\mathbf{x}_i, \theta)} \ ,$$

is very close to that of K-means; the difference is that the assigments of samples to clusters in K-means are discrete.

A main advantage of the mixture model compared to the K-means is that while above the cluster assignments and adaptation formulas were essentially the same, the mixture model can be readily generalized to other kinds of distributions. For instance, the covariance matrix could take on more flexible forms.

A second advantage is that the mixture is now a well-defined distribution, and assuming it is a good model for data, the Bayesian machinery can be used for model selection: choosing the number of clusters or generator distributions. Alternatively, priors can be assigned for a full-Bayesian treatment. We will not discuss these extensions longer here.

2.3 Biclustering

Clustering models group data into groups that are similar over all treatments. This is sensible assuming (i) the treatments form a carefully chosen set that reflects the kinds of functions we are interested in, or (ii) the goal is to separate genes according to differences in their general regulatory mechanisms; then it makes sense simply to collect as varying set of treatments as possible.

Genes may, however, have different functions in different situations, and their regulation may be context-specific as well, meaning that different regulatory factors may be at work in different treatments. *Biclustering* is an extension of clustering where the clustering may be different at different treatments; a cluster may be “active” only in a subset of the conditions. This implies that a gene may belong to different clusters in different treatments; note that this is very different from the “smooth clustering” by mixture models where a gene is assumed to belong to exactly one cluster, we just do not know which one.

Again, there are several different models. We will focus on so-called plaid models [1], which make the additional assumption that data contains an additive mixture of contributions from several clusters. That is, part of the expression of a gene may be due to one regulatory mechanism and part from another. Alternatively, this can be motivated more abstractly as explaining the expression with a set of latent generators, each hopefully corresponding to some as-of-yet unknown cellular function.

Assume that the expression level x_{ij} of gene i in treatment j consists of background activity μ_0 plus a contribution θ_{ijk} if the gene i belongs to the bicluster k and if the bicluster is active in the treatment k. The whole model is

$$x_{ij} = \mu_0 + \sum_k \theta_{ijk} g_{ki} c_{kj} + \varepsilon \ .$$

The activity of the bicluster is modeled by a set of binary-valued variables $c_{kj} = 1$ if cluster k is active in treatment j and $c_{kj} = 0$ otherwise. The binary variables g_{ki} analogously govern whether gene i belongs to cluster k, and ε is normally distributed noise.

The profile of activity of the cluster, given by the θ_{ijk} may depend on the gene or be different for each treatment, or both. This can be formulated as

$$\theta_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk} \ ,$$

where one or both of the variables α and β can be set to zero. The full model is naturally more flexible but also more prone to overfitting.

The original solution to the plaid model was greedy optimization, to successively fit one cluster at a time to maximize the likelihood of the residual remaining from the earlier clusters. The fitting criterion was maximum likelihood, solved by iteratively optimizing one set of parameters at a time. Another proposed solution is full-Bayesian treatment with Gibbs sampling, with computationally convenient (conjugate) priors for the paramaters [2]. The solution was presented only for binarized gene expression data, however.

2.4 Evaluation and use of clusterings

There exist numerous clustering algorithms each having their merits and dismerits in modeling co-expression. The choice of an algorithm is hard both because many have been introduced heuristically, or from different principles, and it is then hard to compare them even in principle. Moreover, the actual problem of modeling co-expression is ill-posed unless more assumptions are made and more data sets are combined. Which aspects of co-expression are important and why specifically would we want to group the genes?

One natural motivation is to model co-expression as merely a proxy to ultimately building mechanistic regulatory models, to be used while we cannot use suitable models for that or do not have the suitable data. Another sensible motivation is the get some intuitive understanding of a new data set or data compendium before further studies. Both are ok but it is good to keep in mind that clustering merely groups genes and does not do miracles.

Irrespective of how and for what the clusters are to be used, it would be crucial to know whether the extracted clusters are real, that is, whether they are properties of the data or artifacts produced by the chosen algorithm. [Unfortunately even this problem is ill-posed, since the (clustering) model gives us our view to the data...]

If we could make a reasonable assumption of what the data should be like if there were no clusters, for instance that it is uniformly distributed, we could measure which model (clusters or no clusters) better fits the data. Unfortunately for high-dimensional real-valued, smallish data sets it is hard to come up with a satisfying definition.

Model-based clusterings can in principle be analyzed with standard probabilistic (Bayesian) techniques, producing confidence intervals. The problem is that we do not usually really think that the data has been generated as a mixture of Gaussians (that is, we know our prior assumptions are wrong), and all such results are conditional on our modeling assumptions.

Alternatively, we can assume that real clusters are robust to small variations in the data, and measure the robustness by randomly re-sampling the data several times and computing clusters from each sample. A problem is that not all interesting clusters such as outlier clusters are necessarily stable in this sense.

Or then we can measure how well the obtained clustering corresponds to known functional classification of the genes, or some other pre-defined information. The obvious problem with this approach is that we are probably applying the clustering to discover something new, instead of replicating the already known information. The clustering will bring new information about new, so-far unlabeled data, though.

Finally, we could generate data where we know the ground truth, run all clustering algorithms on the data and choose the best. The best algorithm would then be run on the gene expression data. The obvious problem here is that while the algorithm was good on the first data there is no general reason why it should be good for the new data as well.

In summary, the clustering solution can be evaluated or validated with several methods, of which none is completely satisfactory. Maybe the best advice is to evaluate using several criteria to make sure the result is sensible, and accept that clustering is an exploratory method meant primarily for producing understandable summaries in data. The summaries are meant for giving sparks for further modeling steps.

3 Differential expression

If it is not possible to remove all uninteresting sources of variation, they can be controlled by making paired measurements where in each pair the uninteresting factors are identical, and only the interesting feature varies. The measurement noise of spotted cDNA arrays stemming from variation of the spots, differenes in binding affinity of the probes, etc. is customarily controlled in this manner. Two samples are hybridized on the same array at the same time, the "case" such as disease tissue and its control ("normal" tissue). The two samples are labeled with different fluorescent dyes, and can be read separately with laser beams having different wavelengths.

In oligonucleotide arrays the measurement process has less noise and it makes sense to measure one sample at a time. But whichever method is used, the sample preparation noise and biological noise remains and needs to be controlled. A simple way is to measure *differential expression*, determine which genes are up- or down-regulated in the case or treatment, compared to a control sample.

We will consider inferring differential expression from a single array with two measurements. Please read T. Jaakkola's lecture notes from the course "Computational functional genomics", lecture 8, available at `http://web.mit.edu/6.874/www/lectures/lecture-8.pdf`.

References

[1] Laura Lazzeroni and Art Owen. Plaid models for gene expression data. *Statistica Sinica*, 12:61–86, 2002.

[2] Qizheng Sheng, Yves Moreau, and Bart De Moor. Biclustering microarray data by Gibbs sampling. *Bioinformatics*, 19(Suppl 2):ii196–ii205, 2003.