CSC 2427: Algorithms in Molecular Biology
Lecture #14

Lecturer: Michael Brudno
Scribe Note: Hyonho Lee
Department of Computer Science
University of Toronto

03 March 2006

Microarrays Revisited

In the last lecture, the guest lecturer, Tim Hughes, talked about microarrays and gene expression. Microarray is basically a two dimensional array, in which each gene or a set of genes are attached to. Using a microarray, we can measure the expression of a certain gene under various circumstances.

Note that a gene is not always expressed. It is sometimes on and sometimes off, depending on various circumstances such as the type of a cell, the external conditions, or the division of a cell. The exception is a *housekeeping gene*, which is always on under any circumstance, since it is needed for the maintenance of the cell.

After doing a series of microarray experiments, we get the result of two dimensional array, where each row represents a gene and each column represents each experiment. For example, we can measure the gene expressions of various types of cells. In these experiments, each column represents the type of a cell: for example, brain cell, liver cell, or cancer cell. We could also measure the gene expressions over time line. For example, the gene expressions of an embryo are changing over time as the embryo develops. In this case, the column of the array represents time line, so we can see in which period each gene is expressed.

Each entry of the array shows the expression of a gene for each experiment. When a gene is expressed, mRNA of a cell binds to the DNA of the microarray. In a microarray, many copies of the same DNAs are attached to each location. So, if there are many mRNAs that bind to the gene, then the microarray shows high level of expression.

There are two main types of microarrays: 1-channel microarray (e.g. Affymetrix) and 2-channel microarray (e.g. cDNA microarray). In 1-channel microarray, we only prepare the test cell. The result shows how much each gene is expressed by the test cell. Usually, greener image shows higher level of gene expression.

In 2-channel microarray, we prepare both the control cell and the test cell. The control cell is usually a mix of all kinds of cell tissues. Then, mRNA of the control sample is dyed green, and the mRNA of the test cell is dyed red. If a gene is more expressed in the control sample than the test cell, then the microarray result shows green. If the gene is more expressed in the test cell than in the control sample, then the microarray shows read. If the gene is equally expressed, then the result is yellow. After getting the result of microarray experiment, we normalize the result to make it comparable.

Using the normalized microarray data, we can make a cluster of genes that have a similar expression pattern or similar gene functions. In another words, we investigate which genes work together. There are many techniques for clustering such as principal component analysis (PCA), independent component analysis (ICA), and Bayesian networks. One way to measure the correlation of two genes is Pearson correlation, which is

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_X S_Y}$$

, where X and Y are the microarray data distribution of two genes, and S_X and S_Y are variations of X and Y, respectively.

Gibbs sampling for motif finding

In the promoter of a gene, there is a *transcription factor binding site* (TFBS), which binds the transcription factors when the gene is expressed. A transcription factor is a protein, and without its binding, RNA polymerase does not transcribe DNA. Since a specific transcription factor binds a specific binding site, it is very useful to find the binding sites in the promoter.

One way to find the binding site is *phylogenetic footprinting*. Since functional sequences are usually well conserved than nonfunctional sequences, we could predict the binding site using footprinting. (This will be covered in the next lecture.) In this lecture, we focus on finding *regulatory motifs*.

Since many genes usually participate in the same process at the same time, many genes tend to be co-expressed. Hence, it is believed that a short motif, which is widespread among many genes, may have an important role to bind the transcription factors.

Regulatory motifs usually have short fixed length. They are repetitive even in a single gene, but very variable. Thus, we want find a pattern rather than a fixed sequence. For example, our target motif would be like

$$\mathrm{G}\ \mathrm{C}\ *\ \overset{\mathrm{T}}{\mathrm{A}}\,\overset{\mathrm{T}}{\mathrm{A}}\,\overset{\mathrm{G}}{\mathrm{T}}\,\mathrm{C}.$$

One solution to find a motif is Gibbs sampling. Gibbs sampling is basically a special case of Monte-Carlo Markov Chain method. Suppose we want to find a motif of length K given t DNA sequences, $X_1, X_2, \ldots, X_t$. Then, the Gibbs sampling algorithm is an iterative algorithm described as follows:

1. After each iteration, we are given t locations $a_1, \ldots, a_t$ for $X_1, \ldots, X_t$, respectively. Let x_j be the substring of X_i starting at a_i.

2. We randomly choose one gene X_i from $X_1, \ldots, X_t$.

3. We calculate a $4 \times K$ position weight matrix (PWM) from the remaining $t-1$ sequences, $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_t$. The each entry $M_{a,b}$ of the PWM indicates the frequency of the nucleotide a at the bth position of x_j's.

4. We also calculate the background probability for each nucleotide. Let

$X_j - x_j$ be the subsequence of X_i after removing x_i. The background probability, B_a, is the proportion of the occurrences of nucleotide a in all t subsequences $X_1 - x_1, \ldots, X_t - x_t$.

5. Now, we are ready to pick a new K long substring from X_i based on the PWM and the background probability. For each K long substring, y_j, starting at j of X_i, we calculate two probabilities, the probability from the PWM and the probability from the background probability. For example, if $y_j = ACGT$, then we calculate $P(y_j|motif) = M_{A,1}M_{C,2}M_{G,3}M_{T,4}$ and $P(y_j|Background) = B_A B_C B_G B_T$. Then, we calculate $\frac{P(y_j|motif)}{P(y_j|Background)}$ for each y_j.

6. We select a position k in X_i based on the odds ratio of $\frac{P(y_j|motif)}{P(y_j|Background)}$. Hence, if a position p has a higher value of $\frac{P(y_j|motif)}{P(y_j|Background)}$, then p is more likely to be chosen as k. We will use $a_1, \ldots, a_{i-1}, k, a_{i+1}, \ldots, a_t$ for the next iteration.

The Gibbs sampling algorithm is very similar to the expectation maximization (EM) algorithm. If we run the Gibbs sampling algorithm infinitely, then it guarantees that we will find the best motif. We normally runs the Gibbs sampling algorithm for a certain number of steps.

In the Gibbs sampling algorithm, we choose a new motif based on the PWM and the background probability. So, if one entry of the PWM is zero, then we never choose a motif that includes the entry. For example, if $M_{A,1} = 0$, then any motif starting with A is never chosen. We do not want this kind of situation, since even though the motif never occurs according to the current PWM, it may still have a chance for next PWM. Thus, we give a very small probability rather than zero when some entry of the PWM is zero. This is called pseudocounts, which is used in the Gibbs sampling algorithm by Lawrence et al [1]. GibbsMotifSampler is a tool using the Gibbs sampling algorithm currently available on the web, and MEME(Multiple EM for Motif Elicitation) [3] is a tool using the EM algorithm.

One common problem of the Gibbs sampling algorithm is that we often encounter the poly-A stretch (AAA…AAA), which is common in genes. To avoid this problem, we employ multi-order Markov model for the background probability. BioProspector [4] uses zero to third-order Markov background models, and BioProspector II uses 7th order Markov background models. CompareProspector [5] uses comparative genomic information and does the Gibbs sampling search with biases towards sequences conserved across species.

References

[1] C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, and J.C. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science Vol. 262*, 1993

[2] M.A. Beer and S. Tavazoie. Predicting gene expression from sequence. *Cell Vol. 117, pp185-198*, 2004

[3] T.L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning, 21(1-2), pp.51-80*, 1995

[4] X. Liu, D.L. Brutlag, and J.S. Liu. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput., pp.127-38*, 2001

[5] Y. Liu, X.S. Liu, L. Wei, R.B. Altman, and S. Batzoglou. Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res., Vol. 14, No. 3., pp.451-458*, 2004