J. Biomedical Science and Engineering, 2010, 3, 977-985 **JBiSE**
doi:10.4236/jbise.2010.310128 Published Online October 2010 (http://www.SciRP.org/journal/jbise/).

Global pattern of pairwise relationship in genetic network

Ao Yuan, Qingqi Yue, Victor Apprey, George E. Bonney

National Human Genome Center, Howard University, Washington DC, USA.
Email: yuanao@hotmail.com; ayuan@howard.edu

Received 6 April 2010; revised 22 April 2010; accepted 10 May 2010.

ABSTRACT

In recent times genetic network analysis has been found to be useful in the study of gene-gene interactions, and the study of gene-gene correlations is a special analysis of the network. There are many methods for this goal. Most of the existing methods model the relationship between each gene and the set of genes under study. These methods work well in applications, but there are often issues such as non-uniqueness of solution and/or computational difficulties, and interpretation of results. Here we study this problem from a different point of view: given a measure of pair wise gene-gene relationship, we use the technique of pattern image restoration to infer the optimal network pair wise relationships. In this method, the solution always exists and is unique, and the results are easy to interpret in the global sense and are computationally simple. The regulatory relationships among the genes are inferred according to the principle that neighboring genes tend to share some common features. The network is updated iteratively until convergence, each iteration monotonously reduces entropy and variance of the network, so the limit network represents the clearest picture of the regulatory relationships among the genes provided by the data and recoverable by the model. The method is illustrated with a simulated data and applied to real data sets.

Keywords: Convergence, Gene-Gene relationship, Neighborhood, Pattern analysis, Relationship measure.

1. INTRODUCTION

A gene regulatory network (also called a GRN or genetic regulatory network) is a collection of DNA segments in a cell which interact with each other (indirectly through their RNA and protein expression products) and with other substances in the cell, thereby governing the rates at which genes in the network are transcribed into mRNA. From methodology point of view, genetic networks are models that, in a simplified way, describe some biological phenomenon from interactions between the genes. They provide a high-level view and disregard most details on how exactly one gene regulates the activity of another. The gene-gene pair wise relationships provide a special insight of the network and are of interest in the study.

Our work is closely related to that of genetic network analysis, and we first give a brief review of the methods. Some methods are deterministic, such as differential (difference) equation models [1-3], which may not be easy to solve nor have unique solutions. Since the genetic network is a complex system, any artificial model can only explain part of its mechanism; the unexplained parts are random noises, so we prefer a stochastic model. Existing stochastic methods for this problem including the linear models [4,5] or generalized linear models [6,7], the Bayesian network [8,9] *etc*. All these methods have their pros and cons, but have the common disadvantage that the solution may not be unique and the results are not easy to interpret. Also, when the network size exceeds that of the data, these methods break down. In genetic work the pair wise regulatory relationships among the genes are important. For such data, it is of interest to investigate the underlying patterns that may have biologic significance, in particular those arising from pair wises regulatory relationships among the genes. Here we study this problem from a different point of view. Given a measure of pair wise gene-gene relationship, we compute the measures from the data, and use the technique of pattern recognition and image restoration to infer the underlying network relationships. The pair wise regulatory relationships among the genes are inferred according to the principle that neighboring genes tend to share some common features, as neighboring genes tend to be co-regulated by some enhancers because of their close proximity [10]. In this method, the solution is unique and computationally simple, the results are easy to interpret and the network can be of any size. In the following we describe our method, study its

basic properties, and illustrate its application. This method is used to reveal the true relationships of structured high dimensional data array [11-14].

2. MATERIALS AND METHODS

The gene expression data are generally time dependent, as in Iyer *et al.* [15]. Let $X_{ij}(t)$ $(i=1,...,m; j=1,...,n; t=1,...k)$ be the observed gene expression response for subject *i*, gene *j* at time *t*. Denote $x_i(t)=(x_{i1}(t),...,x_{in}(t))'$ be the observations across all the genes for subject *i*, and we use $x(t)$ to denote a general sample of the $x_i(t)$'s. Often for this type of data, *m* and *k* are in the low tens, and *n* in the tens to thousands.

The commonly used differential equation model for genetic network analysis is a set of first order homogeneous differential equations with constant coefficients, in the simple case, has the form

$$\frac{dx(t)}{dt}=Wx(t),$$

where $W=(w_{ij})$ is the $n \times n$ matrix of unknown regulatory coefficients to be solved. This type of models and its more specific and complicated variations characterize well the dynamic of the network over time. The base solution of the above equation set is the matrix exponential $e^{tW} := \sum_{r=0}^{\infty} t^r W^r / r! := (v_1(t),...,v_n(t))'$, and the general solution of it has the form $x_i(t)=\sum_{j=1}^{n} c_j v_j(t)$ x_{i} (t), $(i = 1,...,m)$, where the c_j 's are constants to be determined by initial conditions from the data. So there are in total $n^2 + n = n(n + 1)$ coefficients, n^2 of them from *W* and *n* from the c_j 's, to be determined from a total of *mnk* data points. When $mnk < n(n + 1)$ these coefficients can not be determined; when $mnk \geq n(n + 1)$ they may be uniquely or non-uniquely determined, or may still be not determined. For differential (difference) equation models more complicated than this, solutions are more difficult to get.

The commonly used stochastic model is the multivariate linear model

$$x_i(t+1)=Wx_i(t)+\varepsilon_i, \quad E(\varepsilon_i)=0, \quad (i = 1, ..., m; t = 1, ..., k\text{-}1)$$

where $\varepsilon=(\varepsilon_{i1},...,\varepsilon_{in})'$ is the random deviations unexplained by the model. Denote $X(t)=(x_{ij}(t))$, if $X'(t)X(t)$ is non-singular, the least-squares solution of the above model is $W = X'(t + 1) X(t) (X'(t) X(t))^{-1}$, and it may have multiple solutions for different t. For $X'(t_r)X(t_r)$ to be non-singular, one must have $n \leq m$. Even for $n < m$, $X'(t_r)X(t_r)$ may not necessarily be non-singular. This puts an immediate restriction on the size of the network to be analyzed. Also, the solution of the above model may not be unique due to different time points.

For these reasons, we study the problem from a different point of view; by analyze the pair wise gene-gene relationships in the network. In the following we describe our model in which there is always an unique solution, the result is easy to interpret, and there is no restriction on the size of the network. Since the pattern in the genetic network is based on the principle of neighboring similarity, the order of the genes matters in the study, and generally we assume the genes are arranged in their chromosome order.

First we need a measurement for the relationship between any pair of genes, and the network can be represented by the matrix of the pair wise relationships. For large network, linear relationship is not adequate to use, as most of the coefficients will be very small. Also, as mentioned above, such model in this case has no solution because of the small sample size. Pearson's correlation is a good choice for this purpose, other choices including Kendal's tau and Spearman's rho, *etc*. Here we illustrate the method with Pearson's correlation, and our goal is to infer the triangular correlation matrix $R=(r_{ij})_{1\leq i<j\leq n}$ from the observed data, where r_{ij} is the Pearson's correlation coefficient between genes *i* and *j*. As usually the number m of individuals is small (sometimes as few as 2), estimate the correlations using the data at each time point alone is inadequate. So we use all the data to estimate them. An empirical initial version of these correlations are

$$r_{ij}^{(o)}=\frac{1}{mk}\sum_{r=1}^{m}\sum_{s=1}^{k}\frac{(x_{ri}(t)-x_i(t))(x_{sj}(t)-x_j(t))}{\sqrt{Var(x_i(t))Var(x_j(t))}}, \quad ((1\leq i<j\leq n) \qquad (1)$$

where $x_i(t)=\frac{1}{m}\sum_{r=1}^{n}x_{ri}(t)$,

$$Var(x_i(t))=\frac{1}{m}\sum_{r=1}^{n}(x_{ri}(t)-x_i(t))^2, \quad (i = 1,...,n; s=1,...,k).$$

here the $x_{ri}(t)$'s are not i.i.d. over the time *t*'s, and the sample size *mk* is often not large, so the above empirical correlations are very crude evaluations of the true correlations r_{ij} 's. The initial table $R^{(0)}=(r_{ij}^{(0)}:1\leq i<j\leq n)$ is used as the raw data for the next step analysis. For each fixed *i* the observations $x_i(t)$ s at different time conditions reduced the common features in the data, this table is biased as an estimate of *R*. We need to restore their values according to the basic property of the genetic regulatory system. Many reports have shown that nearby genes tend to have similar expression profiles [16-19], thus nearby pairs of genes tend to have similar relationships, and their correlations tend to be close. This

is just the same principle as used in image restoration of data arrays of any size. In the following we use this technique to reduce the bias and improve the estimate of R based on the observation $R^{(0)}$.

Meloche and Zammar [14] considered a method for image restoration of binary data, here we adopt their idea and revise their method to gene expression analysis for continuous data. We assume the following model

$$r_{ij}^{(0)} = r_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0,\sigma^2), (1 \le i < j \le n) \tag{2}$$

for some unknown $\sigma^2 > 0$, where the ε_{ij}'s represent the part of measurements unexplained by the true regulatory relationships in the model. Define the neighbor $R_{ij}^{(0)}$ of $r_{ij}^{(0)}$ to be the collection of the nine immediate members of $R_{ij}^{(0)}$ of $r_{ij}^{(0)} = \{ r_{ab}^{(0)} : | a-i | \le 1, | b-j | \le 1 \}$, which includes $r_{ij}^{(0)}$ itself at the center. For $r_{ij}^{(0)}$'s on the boundary of $R^{(0)}$ the definition is modified accordingly. For example, $R_{1,2}^{(0)}$ and $R_{n-1,n}^{(0)}$ has only three members, $R_{1,j}^{(0)}$ $(3 < j < n\text{-}1)$ has six members, *etc*. Larger neighbors of different shapes can also be considered; here we only illustrate using the above neighbor systems. We assume the $r_{ij}^{(0)}$'s only depend on their neighbors $R_{ij}^{(0)}$'s. The aim is to provide estimates $\hat{r}_{ij}$'s for the true r_{ij}'s based on the records $R^{(0)}$. We assume the estimates have the form for some function $h(\cdot)$ to be specified. The performance of the estimates will be measured by the average conditional mean squared error.

$$r_{ij} = h(R_{ij}^{(0)}), (1 \le i < j \le n), \tag{3}$$

$$\frac{2}{n(n-1)} \sum_{1 \le i < j \le n} E[(\hat{r}_{ij} - r_{ij})^2 \mid R^{(0)}] \tag{4}$$

The optimal set of estimates is the one which minimizes (4). Although r_{ij} is deterministic, we may view it as a realization of the random variable r_{IJ} with (I, J) uniformly distributed over the integer set $S = \{(i, j) : 1 \le i < j \le n\}$. So (4) can be rewritten as

$$EE_{IJ}[(\hat{r}_{IJ} - r_{IJ})^2 \mid R^{(0)}] = E[(\hat{r}_{IJ} - r_{IJ})^2 \mid R^{(0)}]$$
$$= E[(h(R^{(0)}) - r_{IJ})^2 \mid R^{(0)}]$$

Thus by (3), the minimizer of (4) is achieved by

$\hat{r}_{IJ} := h^*(R_{IJ}^{(0)}) = E(r_{IJ} \mid R^{(0)}) = E(r_{IJ} \mid R_{IJ}^{(0)})$, and so

$\hat{r}_{ij} = h^*(R_{ij}^{(0)}) = E(r_{IJ} \mid R^{(0)})$.

To evaluate the above conditional expectation, we need a bit more preparation. Note σ^2 is estimated by

$$\hat{\sigma}^2 = \frac{2}{n(n-1)} \sum_{(i,j) \in S} (r_{ij}^{(0)} - \bar{r})^2, \quad \bar{r} = \frac{2}{n(n-1)} \sum_{(i,j) \in S} r_{ij}^{(0)}.$$

Denote $\varphi(t \mid \sigma^2)$ the normal density function with mean 0 and variance σ^2. Denote S_{ij} as the collection of indices for $R_{ij}^{(0)}$. Given $R_{ij}^{(0)}$, for $(I, J) \in S_{ij}$, view $r_{IJ}^{(0)}$ as a random vector over indices (I,J). We define the conditional distribution of $r_{IJ}^{(0)}$ as

$$P(r_{IJ}^{(0)} = r_{uv}^{(0)} \mid R_{ij}^{(0)}) = \frac{\{\#member \ldots in \ldots R_{ij}^{(0)} = r_{uv}^{(0)}\}}{| S_{ij} |} = \frac{1}{| S_{ij} |}$$

In the above we used the fact that the $r_{uv}^{(0)}$ are continuous random variables, so the collection {members in $R_{ij}^{(0)} = r_{uv}^{(0)}$} = {$r_{uv}^{(0)}$} almost surely. The corresponding conditional probability is defined as

$$P((I,J) = (u,v) \mid R_{ij}^{(0)}) = \frac{\varphi(r_{uv}^{(0)} - r_{ij}^{(0)} \mid \sigma^2) P(r_{uv}^{(0)} \mid R_{ij}^{(0)})}{\sum_{(u,v) \in S_{ij}} \varphi(r_{uv}^{(0)} - r_{ij}^{(0)} \mid \sigma^2) P(r_{uv}^{(0)} \mid R_{ij}^{(0)})}, \text{(u,v)} \in S_{ij}$$

By (2), we deduce $E(r_{IJ} \mid R_{IJ}^{(0)}, (I,J) = (u,v)) = r_{uv}^{(0)}$, so we have

$$\hat{r}_{ij} = E(r_{IJ} \mid R_{ij}^{(0)}) = \sum_{(u,v) \in S_{ij}} E(r_{IJ} \mid R_{ij}^{(0)}, (I,J) = (u,v)) P((I,J) = (u,v) \mid R_{ij}^{(0)})$$

$$= \frac{\sum_{(u,v) \in S_{ij}} r_{uv}^{(0)} \varphi(r_{uv}^{(0)} - r_{ij}^{(0)} \mid \sigma^2) P(r_{uv}^{(0)} \mid R_{ij}^{(0)})}{\sum_{(u,v) \in S_{ij}} \varphi(r_{uv}^{(0)} - r_{ij}^{(0)} \mid \sigma^2) P(r_{uv}^{(0)} \mid R_{ij}^{(0)})}$$

$$= \frac{\sum_{(u,v) \in S_{ij}} r_{uv}^{(0)} \varphi(r_{uv}^{(0)} - r_{ij}^{(0)} \mid \sigma^2)}{\sum_{(u,v) \in S_{ij}} \varphi(r_{uv}^{(0)} - r_{ij}^{(0)} \mid \sigma^2)} \approx \frac{\sum_{(u,v) \in S_{ij}} r_{uv}^{(0)} \varphi(r_{uv}^{(0)} - r_{ij}^{(0)} \mid \hat{\sigma}^2)}{\sum_{(u,v) \in S_{ij}} \varphi(r_{uv}^{(0)} - r_{ij}^{(0)} \mid \hat{\sigma}^2)},$$

$$(i,j) \in S_{ij} \tag{5}$$

The matrix $\hat{R} = (\hat{r}_{ij})$ is our one-step restored estimate of the genetic correlation network R, we also denote it by $R = R^{(1)} = (r_{ij}^{(1)})$.

Denote $F(\cdot)$ the operator given in (5), as $r_{ij}^{(1)} = F(R_{ij}^{(1)})$, and denote $r_{ij}^{(1)} = F(R_{ij}^{(1)})$

$$R^{(1)} = F(R^{(0)}) = E(R \mid R^{(0)}).$$

We view $F(\cdot)$ as a filter for the noises, so $R^{(1)}$ is a smoothed version of $R^{(0)}$. Let $R = r_{IJ}$ be the random variable of the r_{ij}'s over the random index (I,J) and the variation of possible values of the r_{ij}'s, with density $p(\cdot)$, its uncertainty can be characterized by variance and entropy, which is defined as

$$H(p) = -E[\log p(R)] = -\int p(r) \log p(r) dr.$$

It is maximized or most uncertain when R is uniformly distributed, and has smaller value when the distribution of r is more certain. It has some relationship with variance. The former depends on more innate features, such as moments, of the distribution than the latter, which only measures the disparity from the mean. When

$p(\cdot)$ is a normal density with variance σ^2, then $H(p)=1+\sqrt{2\pi\sigma^2}$. For many commonly used parametric distributions, entropy and variance agree with each other, *i.e.* an increase in one of them implies so for the other. But this is not always true and a general closed form relationship between variance and entropy does not exists. Variance is more popular in practice because of its simplicity.

Although generally, in the image restoration context, R is estimated by just applying F once, a natural question is what will happen if we use the operator F repeatedly? *i.e.* let $R^{(k+1)}=(r_{ij}^{(k+1)})=F(R^{(k)})=E(R\mid R^{(k)})$ for $k\geq 0$. To investigate this question, we impose the model

$$r_{ij}^{(k)}=r_{ij}+\varepsilon^{(k)}{}_{ij},\quad \varepsilon^{(k)}{}_{ij}\sim N(0,\sigma^{2(k)}),\ (1\leq i<j\leq n) \tag{6}$$

The estimators $r_{ij}^{(k)}$'s are obtained by minimizing

$$\frac{2}{n(n-1)}\sum_{1\leq i<j\leq n}E[(\hat{r}_{ij}-r_{ij})^2\mid R^{(k)}] \tag{7}$$

and are given by

$$r_{ij}^{(k)}=E(r_{IJ}\mid R_{ij}^{(k)}).$$

similarly $\sigma^{2(k)}$ is estimated by

$$\hat{\sigma}^{2(k)}=\frac{2}{n(n-1)}\sum_{(i,j)\in S}(r_{ij}^{(k)}-\bar{r}^{(k)})^2,$$

$$\bar{r}^{(k)}=\frac{2}{n(n-1)}\sum_{(i,j)\in S}r_{ij}^{(k)}.$$

Since n is usually large, $\hat{\sigma}^{2(k)}$ is a good estimator of $\sigma^{2(k)}$. Corresponding to (5), we have

$$r_{ij}^{(k+1)}=E(r_{IJ}^{(k)}\mid R_{ij}^{(k)})=$$

$$\frac{\sum_{(u,v)\in S_{ij}}r_{uv}^{(k)}\varphi(r_{uv}^{(0)}-r_{ij}^{(0)}\mid\sigma^2)}{\sum_{(u,v)\in S_{ij}}\varphi(r_{uv}^{(0)}-r_{ij}^{(0)}\mid\sigma^2)}\approx\frac{\sum_{(u,v)\in S_{ij}}r_{uv}^{(k)}\varphi(r_{uv}^{(0)}-r_{ij}^{(0)}\mid\hat{\sigma}^{2(k)})}{\sum_{(u,v)\in S_{ij}}\varphi(r_{uv}^{(0)}-r_{ij}^{(0)}\mid\hat{\sigma}^{2(k)})},\quad (i,j)\in S_{ij},\ k\geq 1 \tag{8}$$

In the above we do not replace the $r_{ij}^{(0)}$'s by the $r_{ij}^{(k)}$'s in $\varphi(\cdot\mid\cdot)$ but with $\sigma^{2(0)}$ replaced by the step k estimator $\sigma^{2(k)}$, only for the reason of simplicity in the proof of the Proposition below. Finally, $\sigma^{2(k)}$ is replaced by $\hat{\sigma}^{2(k)}$ in actual computation.

Although few density functions are convex, many of them are log-convex. For example, the normal, exponential (in fact any quadratic exponential families), Gamma, Beta, chisquare, triangle, uniform distributions. But some are not, such as the T and Cauchy distributions. Condition A) does not require all the $p^{(k)}(\cdot)$'s to belong to the same parametric family, nor even to be parametric. Condition B) is satisfied for almost all parametric families as few parametric families require more than the first two moments to determine. The only restriction we make is that all the $p^{(k)}(\cdot)$'s belong to the same parametric family.

View $r^{(k)}$ as a random realization of the $r_{ij}^{(k)}$'s and as of $R^{(0)}$, let $p^{(k)}(\cdot)$ be the density function of $r^{(k)}$. To study the property of the algorithm, we say a non-negative function $f(\cdot)$ is log-convex if $\log f(\cdot)$ is convex, and assume the following conditions

A) $p^{(k)}(\cdot)$ is log-convex for all k.

B) All the $p^{(k)}(\cdot)$'s belong to a parametric family which is determined by the fist two moments.

Our algorithm has the following desirable property (see Appendix for the proof)

Proposition. 1) Assume either A) or B), then

$$H(p^{(k+1)})\leq H(p^{(k)}),\quad k\geq 0.$$

2) $\sigma^{2(k+1)}\leq\sigma^{2(k)}$, $k\geq 0$.

3) As $k\to\infty$, the table $R^{(k)}$ converges in the component wise sense:

$$R^{(k)}\longrightarrow R^*$$

for some stationary array $R^*=R^*(R^{(0)},F)$.

This Proposition tells us that, if the assumption of neighboring similarity is valid for $R^{(0)}$, then the estimates $R^{(k)}$ become more and more clear (less entropy), and more and more accurate as an estimator of R (less variance). So $R^{(*)}$ is the sharpest picture the data $R^{(0)}$ provide and can be restored by the filter F, the innate regulatory relationships among the genes can be recovered by filter F and provided by the data $R^{(0)}$ under the ideal situation of no noise. Intuitively, this picture has some close relationship with the haplotype block structures.

As of small sample size (mk) and large number ($n(n-1)/2$) of parameters, there is no way of talking about the consistency of R to R. So in general $R^{(*)}$ and R may not equal, however our algorithm enable us to do the best effort we can. Convergence of $R^{(k)}$ can be accessed by the distance criteria: for a given > 0 (usually =1/100 or 1/1000)

$$d_1(R^{(k+1)},R^{(k)})=\frac{2}{n(n+1)}\sum_{i<j}|r_{ij}^{(k+1)}-r_{ii}^{(k)}|\leq\varepsilon$$

or

$$d_2(R^{(k+1)},R^{(k)})=\frac{2}{n(n+1)}\Big(\sum_{i<j}(r_{ij}^{(k+1)}-r_{ii}^{(k)})\Big)^{1/2}\leq\varepsilon.$$

Network at each time. We may also investigate the problem at each different time point t. In this case (1) is replaced by

$$r_{ij}^{(o)} = \frac{1}{m}\sum_{r=1}^{m} \frac{(x_{ri}(t)-x_i(t))(x_{rj}(t)-x_j(t))}{\sqrt{Var(x_i(t))Var(x_j(t))}}, (1 \le i < j \le n)$$

where, $x_i(t) = \frac{1}{m}\sum_{r=1}^{m} x_{ri}(t)$,

$Var(x_i(t)) = \frac{1}{m}\sum_{r=1}^{m}(x_{ri}(t)-x_i(t))^2$, $(i = 1,...,n;\ t = 1,...,k)$

and $R^{(0)}(t) = (r_{ij}^{(0)}(t): 1 \le i < j \le n)$ be the corresponding initial table at each t, and the neighborhood for $r_{ij}^{(0)}(t)$ is $R_{ij}^{(0)}(t) = \{r_{ab}^{(0)}: |a-i| \le 1, |b-j| \le 1\}$. In this case (6) is

$r_{ij}^{(k)}(t) = r_{ij}(t) + \varepsilon^{(k)}{}_{ij}(t)$, $\varepsilon^{(k)}{}_{ij}(t) \sim N(0, \sigma^{2(k)}(t))$, $(1 \le i < j \le n)$

and $\hat{r}_{ij}^{(k)}(t) = E(r_{IJ}^{(k)}(t) \mid R_{ij}^{(k)}(t))$.

let $\hat{\sigma}^{2(k)}(t) = \frac{2}{n(n-1)}\sum_{(i,j)\in S}(r_{ij}^{(k)}(t) - \bar{r}^{(k)}(t))^2$,

$\bar{r}^{(k)}(t) = \frac{2}{n(n-1)}\sum_{(i,j)\in S} r_{ij}^{(k)}(t)$.

(8) is now $r_{ij}^{(k+1)}(t) = E(r_{IJ}{}^{(k)}(t) \mid R_{ij}{}^{(k)}) =$

$$\frac{\sum_{(u,v)\in S_{ij}} r_{uv}{}^{(k)}(t)\varphi(r_{uv}^{(0)}(t) - r_{ij}^{(0)}(t) \mid \sigma^2)}{\sum_{(u,v)\in S_{ij}} \varphi(r_{uv}^{(0)}(t) - r_{ij}^{(0)}(t) \mid \sigma^2)}, (i,j) \in S_{ij},$$

$$\approx \frac{\sum_{(u,v)\in S_{ij}} r_{uv}{}^{(k)}(t)\varphi(r_{uv}^{(0)}(t) - r_{ij}^{(0)}(t) \mid \hat{\sigma}^{2(k)}(t))}{\sum_{(u,v)\in S_{ij}} \varphi(r_{uv}^{(0)}(t) - r_{ij}^{(0)}(t) \mid \hat{\sigma}^{2(k)}(t))}$$

$k \ge 1$

The matrix $R^{(k)}(t) = (r^{(k)}{}_{ij}(t))$ is the k-step restored estimate of the genetic correlation network $R(t) = (r_{ij}(t))$ at time t. The proposition is then hold for each fixed t.

3. SIMULATION STUDY

We simulate 40 genes over 12 time conditions at time (hour) points $(t_1,...,t_{12})$ = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12) for 6 individuals by mimicking the setting of the data analyzed in Iyer *et al.* [15]. We simulate the genes from 6 clusters, the numbers of genes in each cluster are given by the vector $(n_1,...,n_6)$ = (8, 6, 4, 4, 6, 12). The baseline values of the gene expressions over time $t \in [0,12]$ in cluster k are generated by functions of the form

$h_k(t) = a_{k1}\sin(t) + a_{k2}\sin(t/2) + a_{k3}\sin(t/3)$, $(k = 1,....,6)$.

Let $h(t)$ be the vector of length 40, with first n_1 components given by $h_1(t)$, second n_2 components by $h_2(t)$,...., last n_6 components by $h_6(t)$. Denote $a_k = (a_{k1}, a_{k2}, a_{k3})$. We arbitrarily choose the a_k's as a_1 = (0.54, –0.18, 1.23), a_2 = (–0.12, –0.25, 0.45), a_3 = (1.0, –0.55, –0.15), a_4 = (–0.32, –0.15, –0.65), a_5 = (0.15, 0.25, 0.35) and a_6 = (–0.52, –0.45, –0.55). First we need to simulate the r_{ij}'s with coordinated patterns. We divide the 40 genes into the 6 clusters, and assume independence among the clusters.

Then for given a covariance matrix $\Omega = \Omega_1 \oplus \cdots \oplus \Omega_6$ we generate the data using this Ω and the time conditions, where Ω_k is the covariance matrix for the genes in cluster k. Directly specifying a high dimensional positive matrix is not easy, we let each Ω_k has the structure $\Omega_k = Q'_k Q_k$, for some Q_k non-singular, so that Ω_k is positive definite. Note that the $Q'_k Q_k$'s may not be correlation matrices, but they are covariance matrices, so is Ω. Let Q_k be upper diagonal with dimension n_k. The non-zero elements of Q_1 are drawn from U(0.5,0.8); those for Q_2 from U(0.2,0.4); those for Q_3 from U(0.2,0.6), those for Q_4 from U(–0.3, –0.1); those for Q_5 from U(0.6,0.9); and those for Q_6 from U(–0.8, –0.6). Then let $\Omega^{1/2} = Q_1 \oplus \cdots \oplus Q_6$. The 6 individuals are i.i.d, so we only need to describe the simulation of observation $x_1 = \{x_{1j}(t): j = 1,...,40; t = 1,...,12\}$ of the first individual. Note for each fixed t, $x_1(t) = (x_{11}(t),...,x_{1,40}(t))$ has covariance matrix Ω. We first generate $y = (y_1,...y_{40})$ with the components i.i.d. N(0,1), then

$$x_1(t) = h(t) + \Omega^{1/2} y + \varepsilon(t)$$

is the desired sample, where for each fixed t, $\varepsilon(t) = (\varepsilon_1(t),...,\varepsilon_{40}(t))$ is the noise, with the $\varepsilon_i(t)$'s i.i.d N(0,1) and independent over t. Convert the covariance matrix $\Omega = (\omega_{ij})$ to a correlation matrix $R = (r_{ij})$ as $r_{ij} = \omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}$ only for $i < j$. Using the data $X = (x_{ij}(t))$, we compute the $R^{(k)}$'s from (8) then use perspective plots to compare the restored correlations after convergence at step k, $R^{(k)}$, the one-step restored $R^{(1)}$, the initial estimated $R^{(0)}$ and the true simulated correlations R.

After computation, the algorithm meets the convergence criterion at iteration 14 with = 10^{-4}. The distances between the observed, first step estimate and last step estimate are: $d_1(R^{(0)}, R)$ = 0.125, $d_1(R^{(1)}, R)$ = 0.108 and $d_1(R^{(14)}, R)$ = 0.094. We see that the estimate after convergence is closest to the true correlations. The results are displayed in **Figure 1**. We only display the correlations r_{ij} for $j > i$. Those values for r_{ij} is 1's, and those for r_{ij} $(i > j)$ are set to zero's, which can be obtained by symmetry.

From this figure we see that the correlations computed from the raw data, panel (b), are very noisy, the true pattern, panel (a), in the network is messed up. The one-step estimate, panel (c), gives some limited sense, while the final estimates, panel (d), recover the true picture with reasonably well. Considering the large number of 40(40+1)/2 = 820 parameters and the small number of 15 individuals on 40 genes, the last step estimates are quite a success. Large number of simulations yield similar results, the convergence criterion is met with 10 to 15 iterations.

Note we only used networks of 40 genes, as large networks are not easy to display graphically. The computations of a network with n genes is in the order $n(n\text{-}1)/2$, so there should be no computational problem for ordinary computer using this method to restore even the whole genome.

4. RESULTS

We use the proposed method to analyze the data with 30 microarray chips from the Stanford microarray database: http:// smd.stanford.edu/cgi-bin/search/QuerySetup.pl. The Category is Normal tissue and the subcategory is PBMC, the following 30 files are the Raw data in the database: 19430.xls, 19438.xls, 19439.xls, 19446.xls, 19447.xls, 19448.xls, 19449.xls, 19450.xls, 19451.xls, 19500.xls, 19505.xls, 19506.xls, 19507.xls, 21407.xls, 21408.xls, 21409.xls, 21410.xls, 21411.xls, 21412.xls, 21413.xls, 21414.xls, 21415.xls, 21416.xls, 21424.xls, 21425.xls, 21426.xls, 21427.xls, 21428.xls, 21429.xls, 21430.xls. The data we used are the overall intensity (mean), the 67th column in the 30 excel files. We choose three subsets of genes on the 30 arrays: set I is genes 0-49, set II is genes 1000-1049 and set III is genes 5000-5049 from the original data set. There are 80 variable for each array. We choose the intensity from normal people for our analysis. The initial correlation coefficients among the genes computed from the raw data in each set, and those estimated after convergence by our algorithm are shown in **Figures 2-4**. Clearly the initial correlations are noisy and difficult to see any patterned

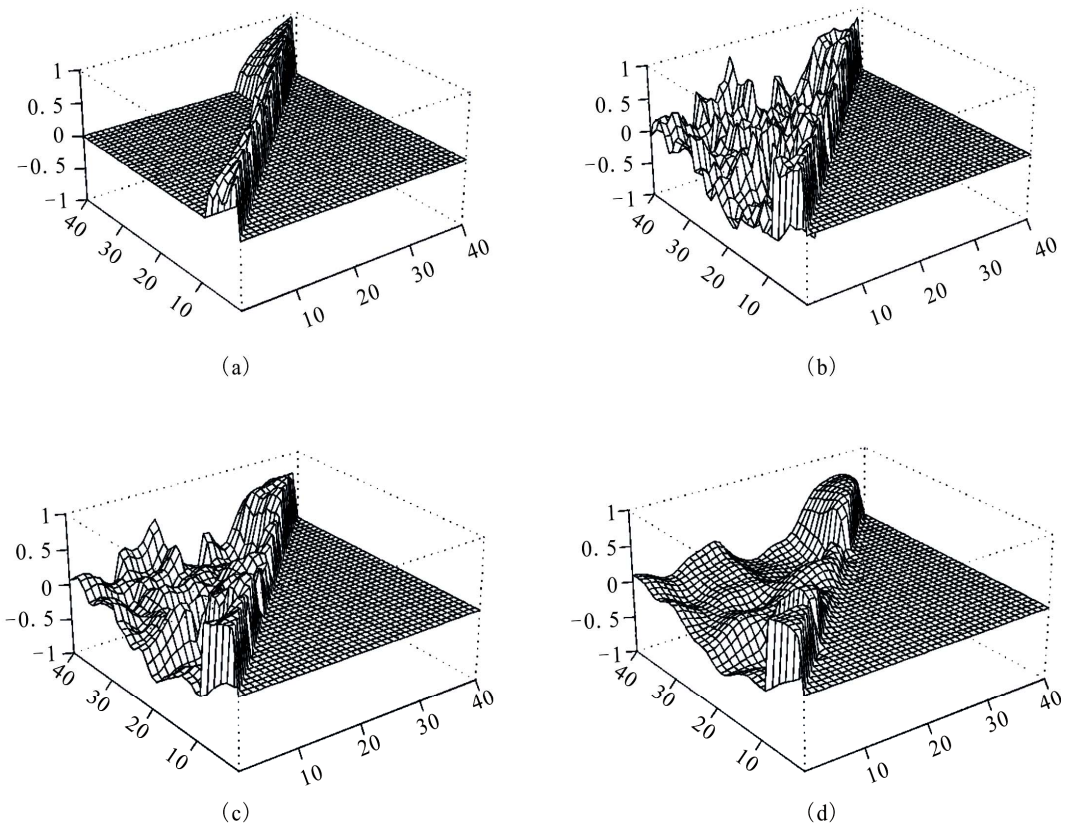

Figure 1. Network Correlations: (a). Simulated R , (b) Initial $R^{(0)}$, (c) One-step Restored $R^{(1)}$, (d) k-step Converged $R^{(k)}$

relationships among the genes. In contrast, the restored pictures are quite clear. For set I, the coefficients are rather homogeneous with values around 0.5, but there is a clear boundary around gene 43, which suggests that most of the genes in this set have similar relationships, or functions. But gene 43 seems to have its own separate mechanism. Genes 38 and 29 also have weak relationships with the other genes. For set II, the relationships among the genes are not so homogeneous. The genes are moderately correlated with coefficients around 0.5, some genes around positions 10, 16, 24, 30, and 38 have weak interactions with the other genes. For set III, there is moderate coordinating pattern among the genes, but three genes, around positions 15, 29, and 40, appears to have relatively independent patterns of regulatory functioning.

5. CONCLUDING REMARKS

We considered a image restoration method for genetic network analysis. This method gives unique solution, the results are easy to interpret and computationally simple. We may implement the genetic distances among the genes into the updating system given in (8). The method is not confined to correlation coefficients among genes, other measures of gene-gene relationships can be considered analogously. Very large networks can be analyzed in principle, the only challenge is how to display the results. We found when the number of genes exceeds 50, the figure is difficult to distinguish visually. The computation for a network of size 40 takes about a couple of minutes using the Splus software. It will be much faster using the C program, and there should be no problem to analyze the whole genome by this method. The only requirement is that the data be arranged in their chromosomal order, otherwise the results may not easy to interpret.

The method can also be used for other analysis purposes and data types, such as cluster analysis. Cluster objects by pattern similarities, *etc*. It can be used to analyze qualitative data type such as haplotype analysis.

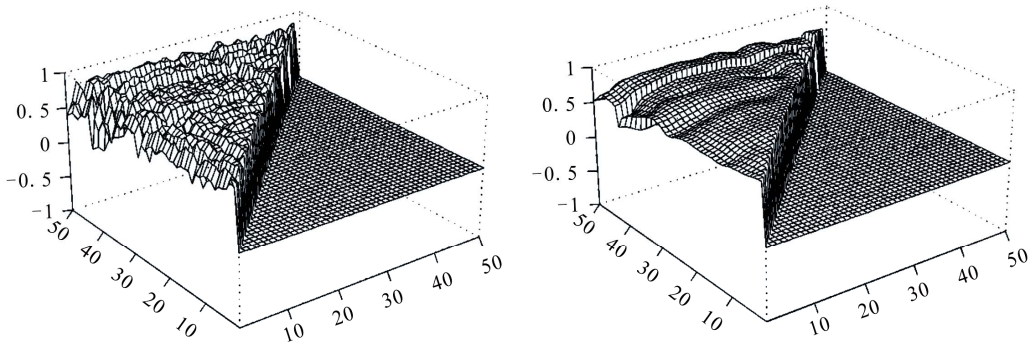

Figure 2. Real data I: initial (left panel) and restored (right panel) correlations.

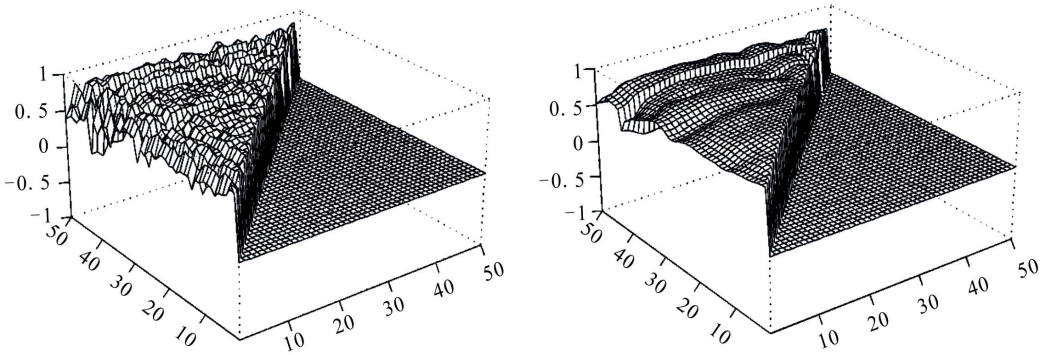

Figure 3. Real data II: initial (left panel) and restored (right panel) correlations.

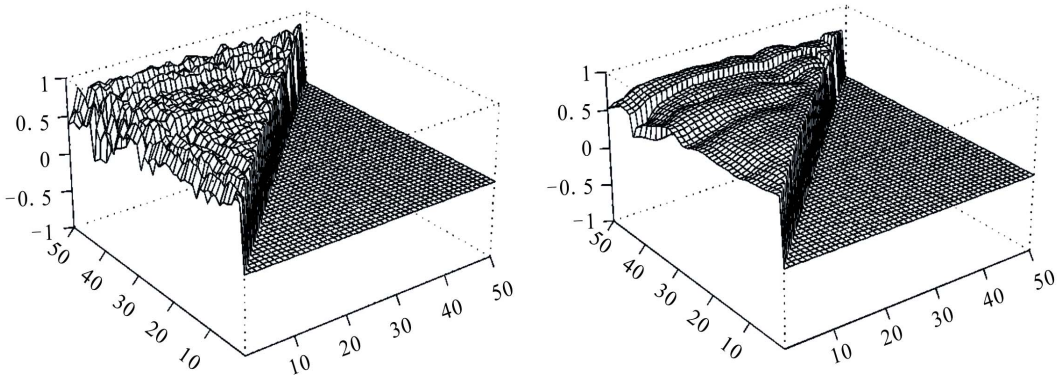

Figure 4. Real data III: initial (left panel) and restored (right panel) correlations.

6. ACKNOWLEDGEMENTS

The research has been supported in part by the National Center for Research Resources at NIH grant 2G12 RR003048.

REFERENCES

[1] Goodwin, B.C. (1965) Oscillatory behavior in enzymatic control processes. In Weber, G., Ed., *Advances in Enzyme Regulation*, Pergamon Press, Oxford, 425-438.

[2] Tyson, J.J. and Othmer, H.G. (1978) The dynamics of feedback cellular control circuits in biochemical pathways. Progress in Biophysics, Academic Press, New York, 1-62.

[3] Reinitz, J., Mjolsness, E. and Sharp, D.H. (1995) Model for cooperative control of positional information in Drosophila by bicoid and maternal hunchback. *Journal of Experimental Zoology*, **271(1)**, 47-56.

[4] Wessels, L.F.A., Van Someren, E.P. and Reinders, M.J.T. (2001) A comparison of genetic network models. *Pacific Symposium on Biocomputing*, **6(4)**, 508-519.

[5] D'haeseleer, P., Liang, S. and Somogyi, R. (1999) Gene expression data analysis and modeling. Pacific Symposium on Biocomputating, Hawaii, USA.

[6] Savageau, M.A. (1976) Biochemical System Analysis: A Study of Function and Design in Molecular Biology. Addison-Wesley, Reading, Massachusetts.

[7] Voit, E.O. (2000). Computational Analysis of Biochemical Systems, Cambridge University Press, Cambridge.

[8] Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000) Using Bayesian network to analyze expression data. *Journal of Computational Biology*, **7(3-4)**, 601-620.

[9] Zhang, B.T. and Hwang, K.B. (2003) Bayesian network classifiers for gene expression analysis. In: Berrar D.P., Dubitzky W. and Granzow M. Ed., *A Practical Approach to Microarray Data Analysis*, Kluwer Academic Publishers, Netherlands, 150-165.

[10] Chen, L. and Zhao, H. (2005) Gene expression analysis reveals that histone deacetylation sites may serve as partitions of chromatin gene expression domains. *BMC Genetics*, **6(1)**, 44.

[11] Owen, A. (1984) A neighborhood based classifier for LANDSAT data. *Canadian Journal of Statistics*, **12(3)**, 191-200.

[12] Ripley, B.D. (1986) Statistics, images, and pattern recognition. *Canadian Journal of Statistics*, **14(2)**, 83-111.

[13] Besag, J. (1986) On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, **48(3)**, 259-302.

[14] Meloche, J. and Zammar, R. (1994) Binary-image restoration. *Canadian Journal of Statistics*, **22(3)**, 335-355.

[15] Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C., Trent, J.M., Staudt, L.M., Hudson J.J., Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D. and Brown, P.O. (1999) The transcriptional program in the response of human fibroblast to serum. *Science*, **283(5398)**, 83-87.

[16] Cohen, B.A., Mitra, R.D., Hughes, J.D. and Church, G.M. (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genetics*, **26(2)**, 183-186.

[17] Caron, H., Schaik, B., Mee, M., Baas, F., Riggins, G., Sluis, P., Hermus, M.C., Asperen, R., Boon, K., Voute, P.A., *et al*. (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, **291(5507)**, 1289-1292.

[18] Lercher, M.J., Urrutia, A.O. and Hurst, L.D. (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genetics*, **31(2)**, 180-183.

[19] Spellman P.T., Rubin G.M. (2002) Evidence for large domains of similarly expressed genes in the Drosophila genome. *Journal of Biology*, **1(1)**, 5.

[20] Ebrahimi, N., Maasoumi, E. and Soofi, E. (1999) Ordering univariate distributions by entropy and variance. *Journal of Econometrics*, **90(2)**, 317-336.

Appendix

Proof of the Proposition. Recall $r^{(k+1)} = r_{IJ}{}^{(k+1)} = r_{IJ}{}^{(k+1)}(R^{(k)})$ is random in (I,J) and $R^{(k)}$; for fixed $(I,J)=(i,j)$, $r_{ij}{}^{(k+1)} = r_{ji}{}^{(k+1)}(R^{(k)})$ is random in $R^{(k)}$; $r_{IJ}{}^{(k+1)} = r_{IJ}{}^{(k+1)}(R^{(k)} \mid R^{(0)})$ is random in (I,J) (discrete); also $r^{(k+1)} = E(r^{(k)} \mid R_{IJ}{}^{(k)}) = E[r_{UV}{}^{(k)} \mid R_{IJ}{}^{(k)}]$ for random index $(I,J) \in S$ and random index $(U,V) \in S_{IJ}$.

1) We first prove the result under condition A). We have

$$E\log p^{(k+1)}(r^{(k+1)}) - E\log p^{(k)}(r^{(k+1)})$$
$$= \int p^{(k+1)}(r)\log\frac{p^{(k+1)}(r)}{p^{(k)}(r)}dr = D(p^{(k+1)} \,||\, p^{(k)}) \geq 0,$$

which is the relative entropy between $p^{(k+1)}(\cdot)$ and $p^{(k)}(\cdot)$. It is known that $D(p^{(k+1)} \,||\, p^{(k)}) \geq 0$ with "=" if and only if $p^{(k+1)}(\cdot) = p^{(k)}(\cdot)$. Note log-convexity of $p^{(k)}(\cdot)$ imply, for each given $R_{IJ}{}^{(k)}$,

$$\log p^{(k)}(E[r^{(k)} \mid R_{IJ}{}^{(k)}]) \geq E[\log p^{(k)}(r^{(k)}) \mid R_{IJ}{}^{(k)}].$$

Thus by the above two inequalities we get

$$H(p^{(k+1)}) = -E\log p^{(k+1)}(r^{(k+1)}) \leq -E\log p^{(k)}(r^{(k+1)})$$
$$= -E\log p^{(k)}(E[r^{(k)} \mid R_{IJ}{}^{(k)}]) \leq E(E[\log p^{(k)}(r^{(k)} \mid R_{IJ}{}^{(k)}])$$
$$= -E\log p^{(k+1)}(r^{(k)}) = H(p^{(k)}).$$

Under condition B), the result in Ebrahimi *et al*. [20] states that entropy and variance agree each other. *i.e*. one increase/decrease implies the other. Now the conclusion is immediate from 2).

2) By the total variance formula, we have

$$\sigma^{(k)} = Var(r_{UV}{}^{(k)})$$
$$= E[Var(r_{UV}{}^{(k)} \mid R_{IJ}{}^{(k)})] + Var(E[r_{UV}{}^{(k)} \mid R_{IJ}{}^{(k)}])$$
$$= E[Var(r_{UV}{}^{(k)} \mid R_{IJ}{}^{(k)})] + Var(r_{IJ}{}^{(k)})$$
$$= E[Var(r_{UV}{}^{(k)} \mid R_{IJ}{}^{(k)})] + \sigma^{(k+1)} \geq \sigma^{(k+1)},$$
$$(k = 0,1,2,\ldots). =$$

3) We only need to prove the convergence of the component $r_{ij}{}^{(k)}$ for any fixed (i,j). In fact from (8), for any integer m and k we have

$$r_{ij}{}^{(k+m)} = \frac{\sum_{(u,v)\in S_{ij}} r_{uv}{}^{(k-1+m)}\varphi(r_{uv}^{(0)} - r_{ij}^{(0)} \mid \sigma^{2(k-1+m)})}{\sum_{(u,v)\in S_{ij}} \varphi(r_{uv}^{(0)} - r_{ij}^{(0)} \mid \sigma^{2(k-1+m)})} := ,$$
$$= F(R_{ij}{}^{(k-1+m)})$$

$(i,j) \in S,\ k,m = 0,1,2\ldots$

Since $\sigma^{(0)} \prec \infty$ and $\sigma^{(k)} \geq 0$, by ii), we have $\sigma^{(k)} \longrightarrow \sigma^*$ for some $0 \leq \sigma^* < \infty$. So if we let

$$\tilde{r}_{ij}{}^{(k+m)} = \frac{\sum_{(u,v)\in S_{ij}} \tilde{r}_{uv}{}^{(k-1+m)}\varphi(\tilde{r}_{uv}^{(0)} - \tilde{r}_{ij}^{(0)} \mid \sigma^{2*})}{\sum_{(u,v)\in S_{ij}} \varphi(\tilde{r}_{uv}^{(0)} - \tilde{r}_{ij}^{(0)} \mid \sigma^{2*})} := ,$$
$$= \tilde{F}(R_{ij}{}^{(k-1+m)})$$

$(i,j) \in S,\ k,m = 0,1,2\ldots$

then $r_{ij}{}^{(k+m)} = \tilde{r}_{ij}{}^{(k+m)} + o(1)$ as $k \to \infty$, thus we only need to prove the convergence of $\{\tilde{r}_{ij}{}^{(k+m)}\}$.

Note

$$\frac{\partial\tilde{F}(R_{ij}{}^{(k-1+m)})}{\partial\tilde{r}_{ij}{}^{(k-1+m)}} = \frac{\varphi(0 \mid \sigma^{2*})}{\sum_{(u,v)\in S_{ij}} \varphi(\tilde{r}_{uv}{}^{(0)} - \tilde{r}_{ij}{}^{(0)} \mid \sigma^{2*})} := C_{ij},$$

We have $0 < C_{ij} < 1$, and for all m,

$$|\tilde{r}_{ij}{}^{(l+1)} - \tilde{r}_{ij}{}^{(l)}| = C_{ij}\,|\tilde{r}_{ij}{}^{(k-1+m)} - \tilde{r}_{ij}{}^{(k-1)}| = \cdots = C^k{}_{ij}\,|\tilde{r}_{ij}{}^{(m)} - \tilde{r}_{ij}{}^{(0)}|$$
$$\leq C^k{}_{ij}\sum_{l=0}^{m-1}|\tilde{r}_{ij}{}^{(l+1)} - \tilde{r}_{ij}{}^{(l)}| \leq C^k{}_{ij}\sum_{l=0}^{m-1}C^l{}_{ij}\,|\tilde{r}_{ij}{}^{(1)} - \tilde{r}_{ij}{}^{(0)}|$$
$$\leq \frac{C_{ij}{}^k}{1-C_{ij}}\,|\tilde{r}_{ij}{}^{(1)} - \tilde{r}_{ij}{}^{(0)}|,$$

thus $\{\tilde{r}_{ij}{}^{(k+m)}\}_{k=1,2,\ldots}$ is a Cauchy sequence, and the convergence follows.

