BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 9 2008, pages 1183–1190
doi:10.1093/bioinformatics/btn098

Systems biology

A pattern recognition approach to infer time-lagged genetic interactions

Cheng-Long Chuang[1,2], Chih-Hung Jen[3], Chung-Ming Chen[1] and Grace S. Shieh[1,2,*]

[1]Institute of Biomedical Engineering, National Taiwan University, Taipei 106, [2]Institute of Statistical Science, Academia Sinica, Taipei 115 and [3]Genome Research Center, National Yang-Ming University, Taipei 112, Taiwan

Received on November 13, 2007; revised on February 4, 2008; accepted on March 6, 2008

Advance Access publication March 12, 2008

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: For any time-course microarray data in which the gene interactions and the associated paired patterns are dependent, the proposed pattern recognition (PARE) approach can infer time-lagged genetic interactions, a challenging task due to the small number of time points and large number of genes. PARE utilizes a non-linear score to identify subclasses of gene pairs with different time lags. In each subclass, PARE extracts non-linear characteristics of paired gene-expression curves and learns weights of the decision score applying an optimization algorithm to microarray gene-expression data (MGED) of some known interactions, from biological experiments or published literature. Namely, PARE integrates both MGED and existing knowledge via machine learning, and subsequently predicts the other genetic interactions in the subclass.

Results: PARE, a time-lagged correlation approach and the latest advance in graphical Gaussian models were applied to predict 112 (132) pairs of TC/TD (transcriptional regulatory) interactions. Checked against qRT-PCR results (published literature), their true positive rates are 73% (77%), 46% (51%), and 52% (59%), respectively. The false positive rates of predicting TC and TD (AT and RT) interactions in the yeast genome are bounded by 13 and 10% (10 and 14%), respectively. Several predicted TC/TD interactions are shown to coincide with existing pathways involving Sgs1, Srs2 and Mus81. This reinforces the possibility of applying genetic interactions to predict pathways of protein complexes. Moreover, some experimentally testable gene interactions involving DNA repair are predicted.

Availability: Supplementary data and PARE software are available at http://www.stat.sinica.edu.tw/~gshieh/pare.htm.

Contact: gshieh@stat.sinica.edu.tw

1 INTRODUCTION

The importance of genetic interactions, which often occur among functionally related genes, lies in the fact that they can predict gene functions (Tong *et al.*, 2004). Understanding genetic interactions in order to unravel the mechanisms of various biological processes in living cells has been a long term endeavor in the field. With the emergence of modern biotechnologies, such as advanced microarray technology, inferring genetic interactions has become feasible.

*To whom correspondence should be addressed.

With the abundant information produced by microarray technology, various approaches have been proposed to infer genetic networks. Most of them may be classified into three classes: graphical models, discrete variable models and continuous variable models. Due to limits of space, we focus on graphical models, which are closest to our approach among the three classes, from the viewpoint of paired similarity and anti-similarity. For the remaining two classes we refer to Shieh *et al.* (2005) for a detailed review.

Graphical models (Whittaker, 1990) depict genetic interactions through directed graphs or digraphs instead of characterizing the interactions quantitatively. Some graphical models simply reveal structural information (Dobra *et al.*, 2004; Kishino and Waddell, 2000; Toh and Horimoto, 2002a; Wu *et al.*, 2003) others annotate the directions and signs of the regulations among genes (Friedman, 2004). Due to their simplicity, graphical models usually require much less data than models used in the other two categories. However, standard theory of graphical Gaussian models (GGMs) requires that the number of data (time points) be greater than the number of genes. In addition, tests for network (model) selection are based on large sample results. Due to these restrictions, the application of GGMs to gene networks has been restricted to small groups of genes (Bay *et al.*, 2002; Waddell and Kishino, 2000; Wang *et al.*, 2003) or a small number of clustered genes (Toh and Horimoto 2002a and b; Wu *et al.*, 2003). Under the limit of sparse (or short time course) microarray data, Bayesian network models have been shown to perform poorly (Husmeier, 2003). On the other hand, the simplicity of GGMs provides opportunities for inferring gene networks. Recently, Wong *et al.* (2005), Dobra *et al.* (2004), and Schäfer and Strimmer (2005) have extended the capacity of GGMs such that large-scale inference on gene networks using small sample data is possible. In particular, Schäfer and Strimmer (2005) employed an empirical Bayes (EB) approach and GGMs (EB-GGMs), in which three partial correlation statistics, an exact test of edge (interaction) inclusion and false-discovery rate (FDR) multiple testing for pairwise significant interactions were introduced. EB-GGMs addressed an important issue in the area of gene networks, namely inferring the gene networks of 3883 genes under the limit of short microarray data.

Besides GGMs, other related works are based on time-lag analysis or correlation coefficient, and are reviewed as follows.

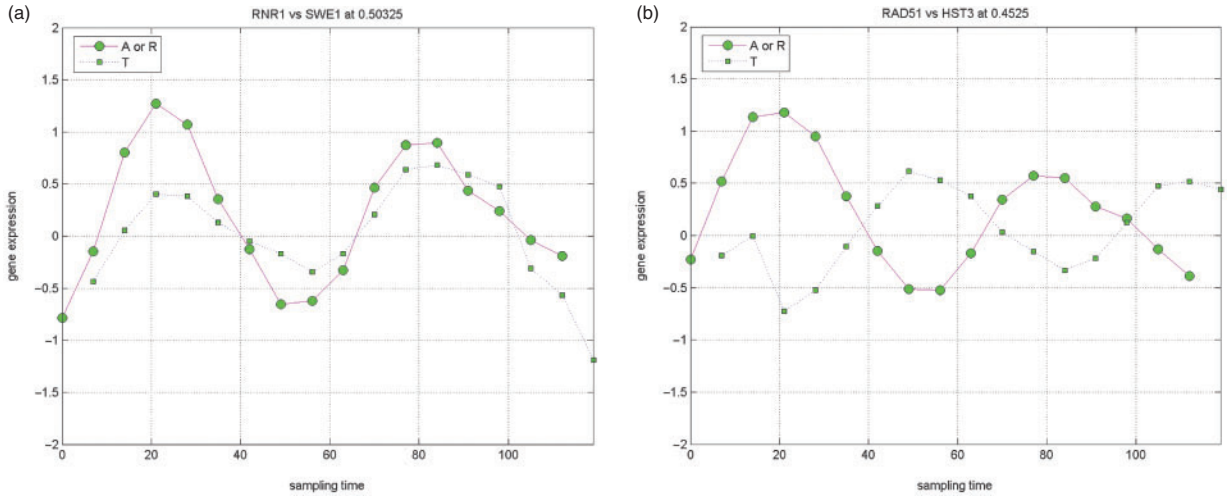

Fig. 1. (**a**) The gene expression pattern of transcriptional diminished (TD) gene pairs RNR1 and SWE1 across time; (**b**) The gene expression pattern of transcriptional compensatory (TC) gene pairs RAD51 and HST3 across time.

Qian *et al*. (2001) utilized the local alignment algorithm to search for time-lagged and inverted correlation. Schmitt *et al*. (2004) employed time-lagged correlation analysis to predict gene-regulatory networks, and the resulting networks allow inference of putative causal relationships among the genes of interest. Zhang and Horvath (2005) proposed robust correlation coefficient to infer co-expression networks. The novelty of Pattern recognition (PARE) lies in two aspects. First, PARE incorporates a non-linear score to identify the time lag of each gene pair, whereas linear correlation coefficients were used in the previous works. Second, PARE integrates existing results from biological experiments or published literature with microarray gene-expression data (MGED) via a machine-learning approach.

Recently, there have been a few studies on transcriptional compensation (TC) interactions (Kafri *et al*., 2005; Lesage *et al*., 2004; Wong and Roth, 2005). Given a synthetic sick or lethal (SSL) or paralog gene pair, following a gene's loss, its compensatory gene's expression increases; this phenomenon is known as TC. Quantitative RT-polymerase chain reaction (qRT-PCR) experiments (in Supplementary Material) showed that besides TC, in some cases following a gene's absence, its SSL partner gene's expression decreased, and we call this phenomenon transcriptional diminishment (TD). While the mechanisms of TC and TD are largely unknown, since genetic interactions provide clues to important biochemical pathways (Shieh *et al*., 2005) inferring such interactions is of interest. In particular, TC and TD interactions between 51 yeast genes that are SSL to SGS1 or RAD27 (Tong *et al*., 2001, 2004) are of interest. Paralogs or redundant genes are called digenic SSL gene pairs, if the combination of two mutants, neither by itself lethal, causes the organism to die or malfunction (Hartman *et al*., 2001; Pan *et al*., 2006; Tong *et al*., 2001). SSL interactions contribute to the robustness of biological pathways. As stated in Tong *et al*. (2004), elements of the genetic networks derived from model organisms are likely to be conserved. Further, for complex heterogeneous human disease syndromes such as glaucoma, type II diabetes, schizophrenia and Alzheimer's disease, a component of the genetic basis of the disease may be similar to the synthetic effects we see in yeast, where multiple pairs of genes have the potential to combine and compromise cellular fitness through a related mechanism. SGS1 (RAD27) has homolog in human cells, including the WRN, BLM and RECQ4 (FEN1 and ERCC5) genes. Mutations in these genes lead to cancer-predisposition syndromes, symptoms resembling premature aging and Cockayne syndrome (Tong *et al*., 2001 and NCBI database).

The proposed approach was implemented on gene pairs which have either direct interactions, for example activator-target (AT), or indirect interactions such as TC. For ease of description, here we utilize AT and repressor-target (RT) to denote TD and TC, respectively. Among qRT-PCR confirmed gene pairs, when the time-course microarray gene expression (Spellman *et al*., 1998) of a target gene T is plotted lag-1 in time behind that of its activator gene (A) or repressor gene (R), in general, AT gene pairs exhibit similar patterns (SP) across time as depicted in Figure 1a. On the other hand, RT gene pairs show complementary gene-expression patterns (one's expression increases while the other's decreases) across time, and we call it complementary pattern (CP) (Figure 1b). These results motivated us to develop an algorithm to learn the pattern of paired gene-expression curves from those known interactions and then the *trained* algorithm can be applied to predict interactions of similar nature. We call this approach PARE algorithm (pronounced as pair).

For any dataset in which the genetic interactions of interest and their corresponding paired gene-expression patterns are significantly associated, PARE can be applied. First, PARE incorporates a non-linear decision score with weights from a pilot study to identify the time lags of gene pairs in a given set, namely identify subclasses with distinct time lags. In each subclass, interactions of some gene pairs are predicted by the non-linear decision score with weights trained by MGED of the other known interactions and some non-interactions

of housekeeping genes (to approximate true pairs with no genetic interactions). Specifically, the non-linear score extracts some characteristics, the first and second derivatives and the enclosed area, of paired gene-expression curves to approximate the non-linear association and dynamics between the curves, and the training applies the particle swarm optimization (PSO) algorithm. Subsequently, PARE predicts the other genetic interactions in the same subclass. Note that genetic interactions cannot be fully mapped by time course MGED. However, this time-lagged approach can partially capture the effects of the protein translation process and post-transcriptional modifications such as phosphorylation and methylation via the time-delayed expression of mRNA.

The rest of this article is organized as follows. Section 2 introduces smoothing of gene expression curves, the procedures to identify the time lags of a given set of gene pairs, the proposed PARE and the optimization algorithm PSO. In Section 3, PARE is applied to real microarray data (Spellman *et al*., 1998) to infer TC interactions. The predictions are checked against 112 pairs of qRT-PCR experimentally confirmed genetic interactions in yeast. False positive rates (FPRs) of PARE for inferring TC (TD) interactions in the yeast genome and 5680 SSL pairs are also reported, respectively. Furthermore, PARE is shown to be able to learn patterns of transcriptional interactions (TIs) from published literature and to predict other TIs adequately. The true-positive rates (TPRs) of PARE on both TC and TIs are compared to those of time-lagged correlation analysis (Schmitt *et al*., 2004) and the latest advance in GGMs (Schäfer and Strimmer, 2005). Besides some novel TC/TD interactions predicted, several predicted TC/TD interactions are shown to coincide with existing pathways. We close with some discussion.

2 SYSTEMS AND METHODS

2.1 Applying mean filter to smooth microarray data

To reduce irrelevant details of expression patterns in MGED so that a trend across time can emerge, we applied a mean filter (Gonzalez and Woods, 2002), in the discrete signal processing area, to smooth microarray data. In general, if the underlying noise follows a Gaussian distribution, a simple mean filter will suffice. A mean filter with kernel size $r \times c$ can be viewed as a window of size $r \times c$ centered at an original datum, and replacing the datum (pixel) with the average of all pixels in the window. Thus the larger the kernel, the smoother the image produced by the filtering. A mean filter with kernel size 1×3 was applied in our study, and its effect is demonstrated in Figure 2. The expression patterns of TC gene pairs (*CSM3* and *HST3*) before and after the filtering are plotted in Figure 2, where the thin solid green (thin dotted blue) line depicts the original expression levels of *CSM3* (*HST3*), while the bold solid green (bold dotted blue) line the expression levels of *CSM3* (*HST3*) after the filtering. Before the filtering, it was hard to observe any pattern from these curves, whereas a CP emerged after the filtering.

2.2 Two paired gene-expression patterns uncovered

The approach in this article was inspired by observing that gene-expression patterns of TC and TD pairs, which were confirmed by qRT-PCR experiments, were of two distinct types. Excluding those pairs amongst which at least one gene had insignificant expression changes across time, when the gene expression curve of a target gene was plotted

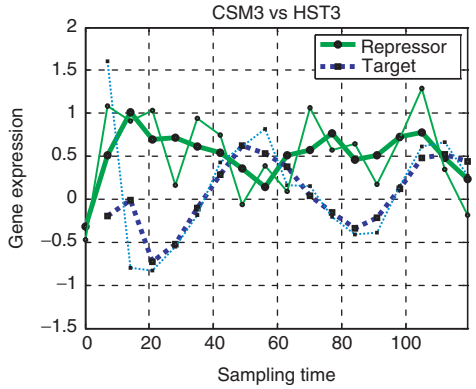

Fig. 2. Paired gene expression patterns before (thin lines) and after (bold lines) the filtering (kernel size 1×3).

lag-1 behind that of its partner gene as depicted in Figure 1a, curves of TD pairs looked similar across time. Whereas gene expression curves of TC pairs, as shown in Figure 1b, exhibited complementary patterns (abbreviated as CPs). A CP is defined as the pattern of paired gene-expression curves where for the majority of time points, a gene's expression increases (decreases) as its associated partner (or target) gene's lag-k expression decreases (increases) across time, for example, the gene expression curves of *RAD51-HST3* in Figure 1b. Note that only significant increases or decreases, e.g. larger than half in $\log_2$ scale, were taken into account. On the other hand, if for the majority of time points, a gene's expression increases (decreases) while its partner (or target) gene's lag-1 expression also increases (decreases), we call this a similar pattern (SP); see the gene-expression curves of *RNR1-SWE1* in Figure 1a for an example. These observations indicate that these paired gene-expression patterns may be a tool to infer genetic interactions, provided that the dependence between each pattern (e.g. CP) and its corresponding interaction (TC) is significant. Thus the underlying assumption of PARE is that paired interactions of interest and their corresponding gene-expression patterns are dependent. This assumption is formally stated in the next section.

2.3 The proposed pattern recognition approach

2.3.1 The condition for applicability of PARE The assumption of PARE corresponds to the null hypothesis that paired interactions (e.g. TC/TD) and paired gene-expression patterns (e.g. CP/SP) are independent, and the alternative hypothesis that paired interactions and the corresponding patterns are not independent. This hypothesis can be tested by Fisher's exact test or χ^2 test, provided that the number of any pattern associated with each type of interaction is less than five or not, using microarray data of known interactions either confirmed by experiments or from published literature. For simplicity, henceforth we only state Fisher's exact test whenever one of these tests is feasible. Any dataset that passes Fisher's exact test can be used to train the parameters in PARE, and then the trained PARE can be applied to predict unknown genetic interactions of similar nature.

For dataset that does not pass Fisher's exact test, the condition can be relaxed. Since only gene pairs with oscillating expression curves may exhibit patterns, genes with flat expression curves can be filtered out by

$$\frac{\max_t(G_i(t))}{\min_t(G_i(t))} > C, \tag{1}$$

where $G_i(t)$ denotes gene i's expression after mean filtered in intensity scale at time t. Our pilot study suggested that the constant C falls in the interval (1.2, 2.0) and can be determined by increasing 0.1 each time from 1.2 until certain proportion of gene pairs formed from

the genes that satisfy Equation (1), and these gene pairs pass Fisher's exact test.

In the following, the 112 pairs of qRT-PCR confirmed TC and TD interactions are utilized to demonstrate this checking. All lag-1 gene pairs did not pass the χ^2 test (=2.33), and its *P*-value 0.13 (not significant at 0.05 level). Thus we relaxed the condition to 30% or more gene pairs formed from genes that passed Equation (1) with $C=1.5$. Note if we relaxed C to 1.4, then there are 66 (58%) gene pairs with lag-1 formed from genes that passed Equation (1).
The 2×2 contingency table of Fisher's exact test is:

	CP	SP	Total
TC	13	9	22
TD	2	11	13
Total	15	20	35

The *P*-value of Fisher's exact test is smaller than 0.02; thus PARE is applicable. The details are in Chisquare-TCTD.pdf of the Supplementary Material. Note that PARE in fact utilizes the maximum of the absolute value of a non-linear score to identify the time lag for each gene pair, and hence the identified time lag will be suitable. This dependence supports the idea that patterns of paired gene-expression curves may be used to distinguish TC pairs from TD pairs. Note that there is a time lag in a (target) gene's expression behind its partner (regulating) gene in our approach. The time lag models a (target) gene that expresses lag-1 behind (is influenced by) the protein of its partner (regulating) gene. As a consequence, the proposed approach may infer the causal relationship of a gene and its partner (target) using time-course microarray data; similar statements can be found in Reis *et al.* (2000), Schmitt *et al.* (2004) and Shieh *et al.* (2004, 2005). Henceforth, the time lag-k for each gene pair is allowed to be different, and the method to identify subclasses of gene pairs with different time lags in a set is explained later in the article.

2.3.2 A score to capture non-linearity in paired gene-expression curves Given any gene pair, their gene-expression curves are likely to show meaningful patterns provided that their time lag is properly identified. In the following, a non-linear score is introduced to capture non-linearity in paired curves, and this score (a function of time lag-k) is used in each subclass with a fixed time lag, say lag-k. For gene-expression curves having SP, their slopes and curvatures tend to share the same signs across most time points. On the other hand, for gene-expression curves with CP, their slopes and curvatures tend to have opposite signs across most time points. To capture characteristics of expression curves with CP and SP, two energy functions based on the products of their first and second derivatives with respect to time are formulated as follows.

$$E_{i,j}^{D1} = \frac{\partial G_i(t')}{\partial t'} \cdot \frac{\partial G_j(t)}{\partial t} \tag{2}$$

and

$$E_{i,j}^{D2} = \frac{\partial^2 G_i(t')}{\partial t'^2} \cdot \frac{\partial^2 G_j(t)}{\partial t^2} \tag{3}$$

where $t'=t+k$, $G_i(t')$ is the lag-k gene-expression level of target gene G_i at time point t' and $G_j(t)$ is the expression level of regulating gene j at time point t. Similar patterns of TD pairs will result in positive values of $E_{i,j}^{D1}$ and $E_{i,j}^{D2}$. On the contrary, $E_{i,j}^{D1}$ and $E_{i,j}^{D2}$ of TC pairs will be negative or small positive values. These two functions are originally from image segmentation literature, but the formulations are novel.

Time-course gene-expression data are discrete in time, so the partial derivatives in Equations (2) and (3) are discretized into

$$\frac{\partial G_i(t)}{\partial t} = \frac{G_i(t+1)-G_i(t)}{\Delta t}$$
$$\frac{\partial^2 G_i(t)}{\partial t^2} = \frac{G_i(t+2)-G_i(t+1)}{\Delta(t+1)\Delta t} - \frac{G_i(t+1)-G_i(t)}{(\Delta t)^2},$$

where $t+1$ denotes the $(t+1)$th time point and $\Delta(t+1)$ is the time interval between time points $t+1$ and $t+2$. When all time intervals are equal, then

$$\frac{\partial^2 G_i(t)}{\partial t^2} = \frac{G_i(t+2)-2G_i(t+1)+G_i(t)}{(\Delta t)^2}$$

Noise in microarray data may contaminate patterns of TC and TD gene pairs, and may lead to high FPRs. To avoid this, a threshold TH is incorporated to filter out the effects of noise as follows.

$$\text{if } \left|G_i'(t)\right| < TH, \text{ set } G_i'(t)=0 \text{ and}$$
$$\text{if } \left|G_i''(t)\right| < TH, \text{ set } G_i''(t)=0.$$

The default value of TH implemented by PARE was 0.1 as commonly used in signal processing literature (Monson, 1996).

We further observed that for gene-expression curves with CP the areas enclosed by $\overline{g_i}(t') = \langle t'-(t'-1), G_i(t')-G_i(t'-1)\rangle$ and $\overline{g_j}(t) = \langle t-(t-1), G_j(t)-G_j(t-1)\rangle$ tended to be large and most of their angles were greater than 90°, whereas gene-expression curves with SP enclosed small areas. Thus we sum the areas enclosed by $\overline{g_i}(t')$ and $\overline{g_j}(t)$ where angles are greater than 90° to form the third function $E_{i,j}^{\text{Area}}$ to distinguish TC from TD pairs, where

$$E_{i,j}^{Area} = \frac{1}{2}\sum_{t\in\Im} \left|\overline{g_i}(t') \times \overline{g_j}(t)\right|,$$
$$\left\{t \in \Im \middle| \theta\left(\overline{g_i}(t'), \overline{g_j}(t)\right) > 90^\circ\right\},$$

and $\theta\left(\overline{g_i}(t'), \overline{g_j}(t)\right)$ is the angle between vectors $\overline{g_i}(t')$ and $\overline{g_j}(t)$. Finally, $E_{i,j}^{D1}$, $E_{i,j}^{D2}$ and $E_{i,j}^{\text{Area}}$ are summed together to form the final decision score

$$E_{i,j} = \alpha \cdot E_{i,j}^{D1} + \beta \cdot E_{i,j}^{D2} - \gamma \cdot E_{i,j}^{\text{Area}},$$

where α, β and γ, are weights to reflect the relative importance of $E_{i,j}^{D1}$, $E_{i,j}^{D2}$ and $E_{i,j}^{\text{Area}}$, respectively.

2.3.3 Identifying subclasses of gene pairs with different time lags The decision score with weights (1, 1, 3.5) was effective at capturing non-linear features of paired gene expression in our pilot study (Chuang *et al.*, 2005); thus this weighted score is used to identify the time lags of gene pairs in any set. Let T_{hcc} and T_{b} denote the time point corresponding to half cell cycle and the last time point to biologists that interactions between genes can occur. The maximum number of time lags $K_0 = \min(T_{\text{hcc}}-1, T_{\text{b}})$ since lagging T_{hcc} reverses an SP to a CP, and a CP to an SP. The procedures to identify the time lags of a set of gene pairs are as follows.

(1) If lag-1 gene pairs, formed from (a) all genes or (b) genes that satisfy Equation (1) with $C \in (1.2, 2.0]$ and pass Fisher's exact test, exceed a certain proportion[1] of the total gene pairs, go to Step 2; otherwise, PARE is not applicable.

[1]Note that this proportion can be as low as 30% since this is a preliminary check for the applicability of PARE. However, the 0.05 level of significance for Fisher's exact test in Steps 3–6 is essential, because this test is to justify that there is indeed a significant association between gene interactions and their corresponding paired patterns. If TPR is not required, then the time lag of each gene pair can be predicted by Steps 1–5. These procedures are demonstrated in the two applications in the RESULTS Section, and have been automated and integrated into the PARE algorithm.

(2) Consider subclasses with all possible combinations of time lags. For example, if $K_0 = 3$, then three subclasses with one time lag, three subclasses with two distinct time lags, and the subclass with all three time lags are considered.

(3) Check whether all subclasses that pass Fisher's exact test include 50% or more total gene pairs, and all paired patterns (e.g. CP and SP) are present. If yes, go to Step 4; otherwise, all gene pairs assume lag 1, output the results and terminate the procedure.

(4) The subclass with distinct time lags identified is the one that yields the maximal overall TPR among all subclasses resulting from Step 3, where the overall TPR is computed from gene pairs in all subclasses that pass Fisher's exact test.

(5) The predicted time lag for each gene pair is the one where the absolute value of the decision score is the largest among all scores computed with time lags in the identified subclasses.

(6) If the identified subclass passes Fisher's exact test or only the subclass with one time lag is left, then assign the time lag for each gene pair accordingly; otherwise, go to Step 3, remove any subclass that does not pass Fisher's test, and reclassify all gene pairs therein. Then iterate Steps 3–6 until Step 3 or 6 is satisfied.

2.3.4 Predicting gene interactions using known interactions In each subclasses with time lag-k, PARE with n-(3-)fold CV can be conducted as follows. For any time lag-k gene pair (G_i, G_i), their interaction is predicted to be TC or TD if the score $E_{i,j}$ computed with lag k, optimized using other lag k gene pairs as training data, is more extreme than the corresponding cutoff values determined from training data in the same subclass. The details of optimization are in the next subsection. The cutoff value for TD (TC) is the minimum (maximum) of the positive (negative) scores that maximizes TPR and true negative rate (TNR) simultaneously in the training set. For instance, for lag-1 gene pairs, we predict the interaction of G_i and G_i is TD or TC, if their score $E_{i,j}$ computed with lag 1 is more extreme than the associated cutoff, e.g. larger than 2.12 or smaller than -1.41, and no interaction if $E_{i,j}$ falls in the interval of the two cutoffs. These two cutoffs are the average of cutoffs in 500 repetitions of n-fold CVs.

2.3.5 Optimizing the weights of the decision score Parameters of some existing pattern recognition algorithms are tuned manually for each set of data in the literature. To automate PARE, the parameters (weights) α, β and γ, are learned from the data. Specifically, these parameters are optimized in the sense that their associated cutoff results in the maximum of δ TPR + $(1-\delta)$ TNR with $\delta = 1/2$ in the training set which includes some qRT-PCR confirmed gene interactions and non-interactions from housekeeping genes. However, optimizing these parameters is complicated, thus PSO is integrated into PARE. If there are two or more sets of parameter values with cutoffs that yield the maximum TPR in the training set, then the set of parameter values whose cutoff is the most extreme is the potentially best solution of α, β and γ according to PSO.

PSO is one of the known efficient optimization methods, and it is a social interactive-based optimization technique proposed by Eberhart and Kennedy (1995).

Assume that a population containing a number of particles, say s, is randomly initialized in the solution space. Each parameter to be optimized corresponds to an axis in the parameter space, and the total number of dimensions is three in this case. We denote the current solution of the ith particle in the jth dimension of the parameter space by $x_{i,j}$. Similarly, we denote the current velocity, best solution of an individual particle and best solution of the entire population by $v_{i,j}$, $bp_{i,j}$, and bg_j, respectively. Each particle searches for the best solution in the unexplored parameter space at time t, and this procedure can be formulated as

$$x_{i,j}(t+1) = x_{i,j}(t) + v_{i,j}(t+1)$$

$$v_{i,j}(t+1) = w \cdot v_{i,j}(t) + c_1 r_{1,i}(t)\left[bp_{i,j}(t) - x_{i,j}(t)\right] + c_2 r_{2,i}(t)\left[bg_j(t) - x_{i,j}(t)\right],$$

where w denotes the inertia weight, which balances the global exploitation and local exploration ability; it is manually chosen from (0,1) according to empirical experience. Often, w is taken to be a constant slightly <1 to reach the global optimum. The acceleration constant c_1 controls how much the particle heads toward its own best solution, and the acceleration constant c_2 controls tendency toward swarm's best ever position; in general, values of c_1 and c_2 are taken to be 1, and $c_1 = c_2 = 1$ were used in our study. Both $r_{1,i}$ and $r_{2,i}$ control the stochastic behavior of the particle movement and they are randomly generated from [0, 1].

After all particles have finished their explorations, solutions obtained by each particle ($bp_{i,j}$) and the population (bg_j) are updated by evaluating the new solution $x_{i,j}$ using training data, e.g. gene interactions confirmed by biological experiments or the literature. If the performance of the new solution $x_{i,j}$ is better than its previous solution $bp_{i,j}$, we update $bp_{i,j}$ to $x_{i,j}$. Moreover, among all those updated $y_{i,j}$, the best solution of the population bg_j is updated to the one that yields the best TPR. If two or more sets of parameters and their corresponding cutoffs result in the same maximum TPR, the best solution of the population bg_j is updated to one that's corresponding cutoff has the largest absolute value. A diagram in PSO.pdf of Supplementary Material illustrates how PSO is incorporated into PARE to optimize parameters.

Let T be the number of time points in a given MGED set, N (N') the number of gene pairs in the training (test) set, M the number of search agents in PSO, and X be the number of generations to achieve the global optimum for PSO. In general, the computational complexity of PARE is in the order of $O(TNMX)$. However, in our applications to genetic networks, since T, M, and X are fixed, the computational complexity of PARE is bounded by max(O(N), O(N')). For details, see PSO.pdf of the Supplementary Material.

3 RESULTS

3.1 MGED

In this section, PARE is applied to the cDNA microarray data in Spellman *et al.* (1998) to infer TC/TD and transcriptional regulatory interactions. For the alpha (Elu) dataset, experiment and control groups were mRNAs extracted from synchronized by alpha factor arrest (elutriation) and non-synchronized yeast cultures, respectively. There were 18 and 14 time points with no replicates in the alpha and Elu datasets, respectively. The red (R) and green (G) fluorescence intensities were measured from the mRNA abundance in the experiment group and control group, respectively. Log ratios of R to G were used to reconstruct the genetic interactions. A full description and complete datasets are available at http://cellcycle-www.stanford.edu.

3.2 Predicting TC and TD genetic interactions

We first applied PARE to infer 112 pairs of TC/TD gene interactions that were confirmed by qRT-PCR experiments. The alpha dataset in Spellman *et al.* (1998) was utilized because 35 gene pairs (31% of this set) passed the χ^2 test with P-value

<0.02, whereas the rest of the datasets did not pass the χ^2 test at the 0.05 significance level.

The procedure to identify subclasses of time lags was implemented as stated in Chisquare-TCTD.pdf of the Supplementary Material. $T_{hcc} = 4$ since a cell cycle consists of nine time points in alpha dataset, and $T_b = 18$ as genetic interactions can last throughout the experiments. Therefore, K_0 equals 3, and the result in Chisqaure-TCTD.pdf show that the subclass with time lag-1 is identified.

Among these gene interactions, $[1-(1/m)]$ were randomly chosen to be a training set to tune the parameters of PARE, and the remaining $1/m$ formed the test set for m-fold cross validation (CV), where $m = 3$ or 112 (leave-one-out CV). One hundred and twenty pairs of housekeeping genes from the SGD website were used to approximate *true negatives* ('pairs with no interactions') in the training and test sets. The cutoff value for TC (TD) using PARE with n-fold CV was −1.41 (2.12), which was the average of 112 sets of cutoffs. Note that 3-fold and n-fold CVs were also performed in the training sets to obtain the averaged accuracy, and if the prediction accuracy of the training set is much larger than that of the test set, then this indicates an over-fitting problem.

Checked against qRT-PCR confirmed interactions and the non-interactions from housekeeping genes, the TPRs (TNRs) of PARE with 3-fold and n-fold CVs were 71% (87%) and 73% (90%), and their CPU time were about 6 and 8 minutes per experiment, respectively. Schäfer and Strimmer (2005) employed an EB-GGMs based on partial correlation to infer genetic networks. EB-GGMs incorporated singular value decomposition and bagging to obtain precise estimates, and employed FDR to reduce falsely predicted interactions. We further applied lagged Pearson correlation (Schmitt *et al.*, 2004) and EB-GMMs to the alpha dataset for a comparison, and the TPRs were 46% and 52%, respectively; the results are summarized in Table 1.

Besides TPR, FPR is also essential to evaluate the performance of an algorithm. According to Wong and Roth (2005), TC and TD interactions are rare. Hence, for simplicity we pretended that there were no TC and TD interactions, and regarded each predicted TC or TD interaction as a false positive to obtain an upper bound of FPR, where FPR is defined as the ratio of the predicted TC (TD) pairs to all gene pairs formed from the yeast genome (6077 genes from http://cellcycle-www.stanford.edu). Among all gene pairs (~37 million pairs), 6.3 million pairs were classified as TCs (their PARE scores < -1.41), hence the FPR for predicting TC (TD) interaction was bounded by 13% (10% with respect to cutoff = 2.12). The FPRs of predicting TC and TD interactions among the 5680 SSL pairs (Wong *et al.*, 2005) were 18% and 12% with respect to the aforementioned cutoffs.

3.3 Predicting transcriptional regulatory interactions

To demonstrate that PARE is applicable to other types of interactions, we applied PARE to infer TIs. One hundred thirty two pairs of TIs, consisting of 76 ATs and 56 RTs, were collected from published literature (Draper *et al.*, 1994; Gancedo, 1998; Mewes *et al.*, 1999). Among the four datasets in Spellman *et al.* (1998), the alpha, cdc28 and Elu datasets passed the χ^2 test that justified the association between each type of interaction and its corresponding pattern. The P-value of the Elu dataset was the most significant among all, thus this set was used. The procedure to identify the time lags of all gene pairs was implemented, and details are in Chisquare-TI.pdf of the Supplementary Material. $T_{hcc} = 6$ since one cell cycle in Elu consists of 13 time points, and $T_b = 4$ since transcriptional regulations occur within 2 h for yeast, and the time interval equals 1/2 h in Elu. Thus K_0 equals 4. The overall TPR of two subclasses (with time lag-1 and lag-2) was the largest among all subclasses that passed Steps 1–3 (details in Chisquare-TI.pdf of Supplementary Material). Therefore, PARE with time lag 1 and lag 2 was carried out. Next, each gene pair was classified to a subclass according to its larger absolute value of the decision score. In each subclass with time lag 1 and lag 2, the χ^2 test (χ^2_1) with Yates' continuity correction was performed, and their values equal 18.3 and 4.63 (P-value = 0.00002 and 0.031), respectively.

Table 1. The prediction results of three algorithms applied to the Alpha data set (18 time points), checked against the 112 pairs of TC/TD interactions confirmed by qRT-PCR

		Training		Test		
		TPR (%)	FPR (%)	TPR (%)	Std (%)	FPR (%)
Lagged Corr.				46		
EB-GGMs				52		
PARE	n-fold	76	20	73		23
	3-fold	78*	18*	71*	3	23*

*The average value of 500 repeated m-fold CVs is reported.

PARE, lagged Pearson correlation (Schmitt *et al.*, 2004) and EB-GMMs (Schäfer and Strimmer, 2005) were applied to infer these 132 TIs and the 120 pairs of *approximate negatives* formed from housekeeping genes. The TPRs (TNRs) of PARE with 3- and n-fold CVs were 74% (85%) and 77%(88%), and their CPU time were about 10 and 12 minutes per experiment, respectively. The cutoffs of PARE for predicting AT (RT) were 5.51 and 6.26 (−6.54 and −5.87) in subclasses with time lag-1 and lag-2, respectively, which were the average of 132 repeated experiments. The TPRs of time-lagged correlation and EB-GMMs were 51 and 59%, respectively; the results are summarized in Table 2. Similar to the TC/TD application, the FPRs for genome-wide AT and RT prediction were computed, and they were bounded by 10 and 14%, respectively. The list of 132 TIs, their PARE scores, molecular functions, and references are summarized in TI.pdf of Supplementary Material. The high TPRs and reasonable FPRs of PARE in inferring both TIs and TC/TD interactions show that a machine-learning approach can be powerful if information from biological experiments or published literature is properly integrated with MGED.

3.4 Several predicted TC/TD interactions coinciding with existing pathways

PARE with n-fold CV successfully uncovered TC/TD interactions among genes involved in different biological functions,

Table 2. The prediction results of three algorithms applied to Elu data set (14 time points), checked against the 132 transcriptional interactions from literatures

		Training		Test		
		TPR (%)	FPR (%)	TPR (%)	Std (%)	FPR (%)
Lagged Corr.				51		
EB-GGMs				59		
PARE	*n*-fold	79	16	77		17
	3-fold	81*	16*	74*	3	19*

*The average value of 500 repeated *m*-fold CVs is reported.

such as DNA replication (e.g. SRS2, SLX1 and MUS81), maintenance of chromosome complex (SLX1 and MUS81), and DNA repair (e.g. RAD51, RAD52 and MUS81), checkpoint arrest (e.g. RAD9) and chromosome segregation (e.g. CSM3). Besides some novel predicted TC/TD interactions, the following are consistent with existing pathways confirmed by experiments (via iHOP database, Hoffmann and Valencia, 2004). Defects in RAD51 and other homologous recombination genes suppressed synthetic lethality/sickness of the double mutant *sgs1Δ srs2Δ*. Slx1-Slx4 was found to be a second structure specific endonuclease functionally redundant with Sgs1-Top3 in Fricke and Brill (2003). Sgs1/Top3/Rmi1 and Mus81/Mms4 complex are involved in both double-strand break repair and homologous recombination (Fabre *et al*., 2002). This indicates that Sgs1/Top3/Rmi1 and Mus81/Mms4 are alternate pathways to resolve recombination intermediates. Onoda *et al*. (2001) identified that Sgs1 participated in a RAD52-dependent recombinational pathway. Ooi *et al*. (2003) found that Rad9 and Sgs1 interact genetically and possibly physically. Cells lacking Sgs1 frequently arrest as large-budded cells with a single nucleus in the mother cell, or 'stuck' between mother and daughter cells which resulted in missegregation during mitosis (Lo *et al*., 2006; McVey *et al*., 2001), whereas Csm3 is required for DNA replication checkpoint and accurate chromosome segregation. Srs2 and Cms3 were involved in promoting post-replication repair and replication fork restart (Xu *et al*., 2004). Srs2 and Mus81 were identified acting in the same pathway for repair of spontaneous DNA lesion (Schmidt and Kolodner, 2004) and preventing formation of toxic recombination (Fabre *et al*., 2002). Slx1 and Mus81 assembled into two structure-specific endonucleases were required in the Brc1-mediated rescue of Smc5/6 deficiency (Lee *et al*., 2007). Rad51, Rad9, Rad52 and Mus81 participated in repairing endogenous Apurinic/apyrimidinic (AP) sites that are one of the most frequent endogenous lesions in DNA (Boiteux and Guillet, 2006). These consistencies reinforce the possibility of applying genetic interaction results to predict pathways of protein complexes (Collins *et al*., 2007).

4 DISCUSSION

The proposed PARE algorithm first identifies time lags for all gene pairs in a given set, namely it determines how many subclasses are there with different time lags. Next, in each subclass, unknown interactions of some gene pairs can be predicted by the non-linear decision score with weights learned from MGED of the other known interactions. TPRs of PARE applied to the datasets in Spellman *et al*. (1998) range from 71 to 77%, which are significantly higher than a time-lagged correlation approach and the latest advance in GGMs. PARE, being a machine-learning approach, has the advantage of naturally integrating MGED with information from biological experiments or published literature. The results in Section 3 indicate that PARE may be an exploratory tool to infer gene networks with limited microarray data. Furthermore, several predicted TC/TD interactions are shown to coincide with existing pathways, and these consistencies indicate that PARE has potential to predict pathways too.

As suggested by a reviewer, we checked both applications, and there were cases in which the time-lagged correlation approach and/or EB-GGMs predicted correctly but PARE failed. In view of this, a promising research direction would be to combine linear and partial correlations with PARE, and also employ FDR to test predicted interactions further. Recently incorporating various types of genomic data, e.g. motif information, ChIP-chip data and microarray data, to predict transcriptional modules have been explored (Lemmens *et al*., 2006; Tsai *et al*., 2005, 2006). However, integrating various types of data for reliable prediction of complex genetic networks remains a challenging topic. Given the high TPRs and reasonable FPRs, PARE may be a feasible tool to integrate other types of genomics data with MGED, and we leave this for future research.

ACKNOWLEDGEMENTS

We wish to thank Dr Ting-Fang Wang at the Institute of Biological Chemistry, Academia Sinica for providing the qRT-PCR results, Dr Shang-Kai Tai and Mr. Chia-Chang Wang for collecting TIs, Drs Yuan-Chin Chang and Ker-Chau Li for very helpful discussions, and former Vice President Prof. Ming-Chiao Lai for encouragement and support. We are grateful to A.E. and two anonymous referees for their invaluable suggestions to improve the paper. This research was supported by NSC grant 95-2118-E-002-029-MY2 (for Chen), and NSC grant 94-2118-M-001-026 and Academia Sinica 95-TP grant 23-33 (for Shieh).

REFERENCES

Bay,S.D. *et al*. (2002) Revising regulatory networks: from expression data to linear causal models. *J. Biomed. Inform*., **35**, 289–297.

Boiteux,S. and Guillet,M. (2006) Use of yeast for detection of endogenous abasic lesions, their source, and their repair. *Methods Enzymol*., **408**, 79–91.

Chuang,C.L. *et al*. (2005) A pattern recognition approach to infer genetic networks. *Technical Report C2005-05*, Institute of Statistical Science, Academia Sinica, Taiwan.

Collins,S.R. *et al*. (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, **446**, 806–810. February 21, 2007 [Epub ahead of print].

Dobra,A. *et al*. (2004) Sparse graphical models for exploring gene expression data. *J. Multiv. Anal*., **90**, 196–212.

Draper,M.P. *et al*. (1994) CCR4 is a glucose-regulated transcription factor whose leucine-rich repeat binds several proteins important for placing CCR4 in its proper promoter context. *Mol. Cell. Biol*., **14**, 4522–4531.

Eberhart,R.C. and Kennedy,J. (1995) A new optimizer using particle swarm theory. In *Proceeding of 6th International Symposium. Micro Machine and Human Science*. IEEE service center, Piscataway, NJ, Nagoya, pp. 39–43.

Fabre, F., Chan, A., Heyer, W.D. and Gangloff, S. (2002) Alternate pathways involving Sgs1/Top3, Mus81/Mms4, and Srs2 prevent formation of toxic recombination intermediates from single-stranded gaps created by DNA replication. *Proc. Natl Acad. Sci. USA*, **99**, 16887–16892.

Fricke, W.M. and Brill, S.J. (2003) Slx1-Slx4 is a second structure-specific endonuclease functionally redundant with Sgs1-Top3. *Genes Dev*., **17**, 1768-1778.

Friedman,N. (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.

Gancedo,J.M. (1998) Yeast carbon catabolite repression. *Microbiol. Mol. Biol. Rev*., **62**, 334–361.

Gonzalez,R.C. and Woods,R.E. (2002) *Digital Image Processing*. 2nd edn. Prentice Hall, Upper Saddle River, NJ.

Hartman,J.L. *et al*. (2001) Principles for the buffering of genetic variation. *Science*, **291**, 1001–1004.

Hoffmann,R. and Valencia,A. (2004) A gene network for navigating the literature. *Nat. Genet*., **36**, 664.

Husmeier,D. (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, **19**, 2271–2282.

Kafri,R.A. *et al*. (2005) Transcription control reprogramming in genetic backup circuits. *Nat. Genet*., **37**, 295–299.

Kishino,H. and Waddell,P.J. (2000) Correspondence analysis of genes and tissue types and finding genetic links from microarray data. *Genome Inform*., **11**, 83–95.

Lee,K.M. *et al*. (2007) Brc1-mediated rescue of Smc5/6 deficiency: requirement for multiple nucleases and a novel Rad18 function. *Genetics*, **175**, 1585–1595.

Lemmens,K. *et al*. (2006) Inferring transcriptional modules from ChIP-chip, motif and mircoarray data. *Genome Biol*., **7**, R37.

Lesage,G. *et al*. (2004) Analysis of beta-1,3-glucan assembly in *Saccharomyces cerevisiae* using a synthetic interaction network and altered sensitivity to caspofungin. *Genetics*, **167**, 35–49.

Lo,Y.C. *et al*. (2006) Sgs1 regulates gene conversion tract lengths and crossovers independently of its helicase activity. *Mol. Cell. Biol*., **26**, 4086–4094.

McVey,M. *et al*. (2001) The short life span of Saccharomyces cerevisiae sgs1 and srs2 mutants is a composite of normal aging processes and mitotic arrest due to defective recombination. *Genetics*, **157**, 1531–1542.

Mewes,H.W. *et al*. (1999) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*., **27**, 44–48.

Monson,H.H. (1996) *Statistical Digital Signal Processing and Modeling*. Wiley, New york.

Onoda,F. *et al*. (2001) Involvement of SGS1 in DNA damage-induced heteroallelic recombination that requires RAD52 in Saccharomyces cerevisiae. *Mol. Gen. Genet*., **264**, 702–708.

Ooi,S.L. *et al*. (2003) DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray. *Nat. Genet*., **35**, 277–286.

Pan,X. *et al*. (2006) A DNA integrity network in the yeast *Saccharomyces cerevisiae*. *Cell*, **124**, 1069–1081.

Qian,J. *et al*. (2001) Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles indentified new, biologically relevant interactions. *J. Mol. Biol*., **314**, 1053–1066.

Reis,B.Y. *et al*. (2000). Approaching causality: discovering time-lag correlations in genetic expression data with static and dynamic relevance networks. In *Proceedings of RECOMB 2000*, Tokyo, Japan. p. 5.

Schäfer,J. and Strimmer,K. (2005) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**, 754–764.

Schmidt,K.H. and Kolodner,R.D. (2004) Requirement of Rrm3 helicase for repair of spontaneous DNA lesions in cells lacking Srs2 or Sgs1 helicase. *Mol. Cell. Biol*., **24**, 3213–3226.

Schmitt,W.A. *et al*. (2004) Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. *Genome Res*., **14**, 1654–1663.

Shieh,G.S. *et al*. (2004) A regression approach to reconstruct gene networks. In *Proceedings of 2004 Taipei Sympisium on Statistical Genome*, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, pp. 357–370.

Shieh,G.S. *et al*. (2005). A stepwise structural equation modeling algorithm to reconstruct genetic networks. *Technical Report C2005-04,* Institute of Statistical Science, Academia Sinica, Taiwan.

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Sarcharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Toh,H. and Horimoto,K. (2002a) Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics*, **18**, 287–297.

Toh,H. and Horimoto,K. (2002b) System for automatically inferring a genetic network from expression profiles. *J. Biol. Phys*., **28**, 449–464.

Tong,A.H. *et al*. (2001) Systematic genetic analysis with ordered arrays of Yeast deletion mutants. *Science*, **294**, 2364–2366.

Tong,A.H. *et al*. (2004) Global mapping of the Yeast genetic interaction network. *Science*, **303**, 808–813.

Tsai,H.K. *et al*. (2005) Statistical methods for identifying yeast cell cycle transcription factors. *Proc. Natl Acad. Sci*., **12**, 13532–13537.

Tsai,H.K. *et al*. (2006) Method for identifying transcription factor binding sites in yeast. *Bioinformatics*, **22**, 1675–1681.

Waddell,P.J. and Kishino,H. (2000) Cluster inferences methods and graphical models evaluated on NCI60 microarray gene expression data. *Genome Inform*., **11**, 129–140.

Whittaker,J. (1990) *Graphical Models in Applied Multivariate Statistics*. Wiley, NY.

Wang,F. *et al*. (2003) Efficient estimation of covariance selection models. *Biometrika*, **90**, 809–830.

Wong,S.L. and Roth,F.P. (2005) Transcriptional compensation for gene loss plays a minor role in maintaining genetic robustness in saccharomyces cerevisiae. *Genetics*, **171**, 829–833.

Wong,S.L. *et al*. (2005) Discovering functional relationships: biochemistry versus genetics. *Trends Genet*., **21**, 424–427.

Wu,X. *et al*. (2003) Interactive analysis of gene interactions using graphical Gaussian model. In *Proceedings of the ACM SIGKDD Workshop on Data Mining in Bioinformatics*. Vol. 3. ACM Press, Washington, DC, USA, pp. 63–69.

Zhang,B. and Horvath,S. (2005) A general framework for weighted gene co-expression networks analysis. *Stat. Appl. Genet. Mol. Biol*., **4**, Article 17.

Xu,H. *et al*. (2004) Mrc1 is required for sister chromatid cohesion to aid in recombination repair of spontaneous damage. *Mol. Cell. Biol*., **24**, 7082–7090.