Inference of Regulatory Gene Interactions from Expression Data Using Three-Way Mutual Information

John Watkinson,[a,c] Kuo-ching Liang,[a] Xiadong Wang,[a] Tian Zheng,[b] and Dimitris Anastassiou[a,c]

[a]Department of Electrical Engineering, [b]Department of Statistics, [c]Center for Computational Biology and Bioinformatics, Columbia University, New York, New York, USA

This paper describes the technique designated best performer in the 2nd conference on Dialogue for Reverse Engineering Assessments and Methods (DREAM2) Challenge 5 (unsigned genome-scale network prediction from blinded microarray data). Existing algorithms use the pairwise correlations of the expression levels of genes, which provide valuable but insufficient information for the inference of regulatory interactions. Here we present a computational approach based on the recently developed context likelihood of related (CLR) algorithm, extracting additional complementary information using the information theoretic measure of synergy and assigning a score to each ordered pair of genes measuring the degree of confidence that the first gene regulates the second. When tested on a set of publicly available *Escherichia coli* gene-expression data with known assumed ground truth, the synergy augmented CLR (SA-CLR) algorithm had significantly improved prediction performance when compared to CLR. There is also enhanced potential for biological discovery as a result of the identification of the most likely synergistic partner genes involved in the interactions.

Key words: **regulatory networks; computational methods**

Introduction

Several techniques for gene interaction network inference from microarray data have been proposed and used successfully,[1] such as those based on pairwise mutual information,[2,3] Bayesian networks,[4,5] regression techniques,[6] and graphical Gaussian models.[7,8] There is not a universally accepted definition of the word "interaction," but in this paper we address the more clearly defined special case of gene regulatory networks using directed graphs in which nodes represent genes and directed edges imply that the product of the gene represented by the head node participates in the regulatory mechanism of the gene represented by the tail node. Bayesian networks are, by design, suited for such inference, but a fundamental limitation stems from the requirement that the topology of the network graphs makes use of heuristic or greedy algorithms.

Information theoretic tools are useful for identifying statistical dependencies between the expression levels, indicating functional relationship of the corresponding genes. Specifically, use of the pairwise mutual information[9] as a measure of correlation between two genes, when accompanied by filtering to reduce inference of interactions via intermediaries, has been demonstrated to be a powerful tool (ARACNe) for the reverse engineering of cellular networks.[10,11] More recently, it was proposed[12] that the values of the pairwise mutual information undergo adaptive background correction to eliminate false correlations, and this concept led to the context likelihood of

Address for correspondence: Dimitris Anastassiou, Department of Electrical Engineering, Columbia University, 1312 S.W. Mudd Building, Mail Code 4712, 500 West 120th Street, New York, NY 10027. anastas@ee.columbia.edu

The Challenges of Systems Biology: Ann. N.Y. Acad. Sci. 1158: 302–313 (2009).
doi: 10.1111/j.1749-6632.2008.03757.x

relatedness (CLR) algorithm, which had outstanding transcriptional interaction prediction performance when tested on a set of *Escherichia coli* expression data.

When the expression levels of two genes exhibit high correlation as evidenced from measurements under mixed and diverse conditions, this may indicate that one of the genes participates in the regulatory mechanism of the other, but not necessarily. For example, the two genes can be, at least partly, coregulated by a shared mechanism. Conversely, even if a gene is a direct regulator of another gene, this does not necessarily imply that the expression levels of the two genes will be significantly correlated, as this regulation may only manifest itself cooperatively with other regulating genes. Indeed, weak pairwise correlations have been shown to occasionally imply strongly correlated network states in other contexts, such as neural populations.[13] Therefore, to infer whether a gene regulates another gene from microarray data, it is important to consider the influence of additional genes on the potential interaction. One such proposed approach[14] models the dependency of transcriptional interactions on the availability of the products of "modulator" genes by contrasting the pairwise correlations of the expression levels of the two interacting genes observed under the highest and the lowest expression levels of such potential modulators.

Here we present an algorithm that uses the information theoretic measure of synergy,[15] leading to a novel methodology for developing a directed graph depicting inferred regulatory interactions without requiring prior biological knowledge or discretization of the expression values. The concept of synergy can be used for the inference of gene interactions with respect to a particular phenotype such as cancer.[16] For the current application, we apply the concept without reference to a phenotype, resulting in the assignment, to each directed pair of genes, of a numerical score indicating the degree of confidence that the product of the gene represented by the head node is one of the regulators

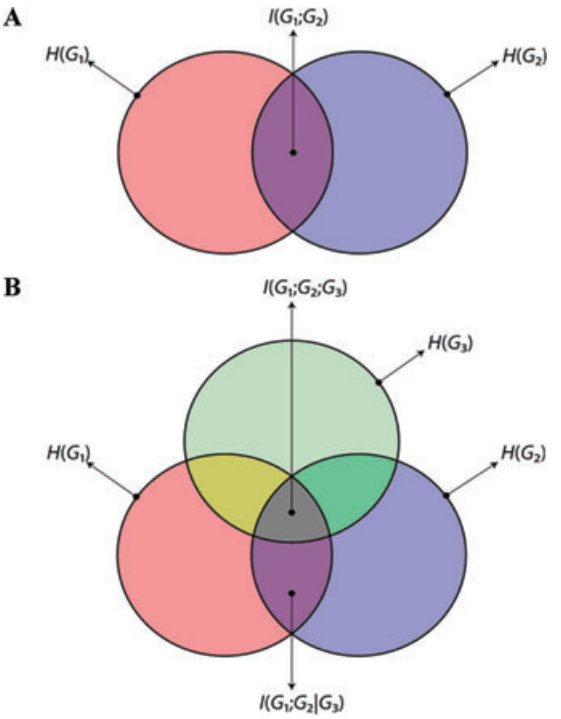

FIGURE 1. Venn diagram representation of mutual information. (**A**) The pairwise mutual information $I(G_1; G_2)$ corresponds to the intersection of the two circles and is always nonnegative. (**B**) The three-way mutual information $I(G_1; G_2; G_3)$ corresponds to the intersection of the three circles, but it is not always nonnegative. If it is negative, then there is no Venn diagram representation possible and $-I(G_1; G_2; G_3)$ is equal to the synergy of each pair of two genes with respect to the third gene.

of the gene represented by the tail node. The computational aspects are described in Methods, and our results using a synthetic and a real expression dataset with known ground truth are described in Results and Discussion below, making use of the following concepts.

Assuming that the continuous expression levels of two genes G_1 and G_2 are governed by a joint probability density p_{12} with corresponding marginals p_1 and p_2 and using simplified notation, the mutual information $I(G_1, G_2)$ is defined as the expected value $E\{\log \frac{p_{12}}{p_1 p_2}\}$. It is a non-negative quantity representing the information that is common to the two variables and can be depicted in a Venn diagram (Fig. 1A) as corresponding to the intersection of the

information in G_1 with the information in G_2, where the area of each region in the diagram measures the corresponding entropy.[9] The pairwise mutual information has successfully been used as a general measure of the correlation between two random variables.

Generalizing this concept and notation to include three genes G_1, G_2, and G_3, we define the symmetric quantity $I(G_1; G_2; G_3)$ as the expected value $E\{\log \frac{p_{12}p_{23}p_{13}}{p_1p_2p_3p_{123}}\}$. This "three-way mutual information" can be seen[9,15] as representing the information that is common to the three variables compatible with a Venn diagram representation (Fig. 1B) as the intersection of three regions. The three-way mutual information is not necessarily non-negative. When it is negative, then the positive quantity $-I(G_1; G_2; G_3)$ is equal to the *synergy* of two of these variables with respect to the third, where the synergy of two variables G_1, G_2 with respect to a third variable G_3 is[15] equal to $I(G_1, G_2; G_3) - [\, I(G_1\,; G_3) + I(G_2\,; G_3)\,]$, that is, the part of the association of a pair of genes G_1, G_2 with gene G_3 that is purely due to a synergistic cooperation between genes G_1 and G_2 (the "whole" minus the sum of the "parts"). In that case, the positive quantity $-I(G_1; G_2; G_3)$ can be seen as measuring the "entanglement" connecting the three genes, and there is no possible Venn diagram representation. One could consider the analogy that the intersection of the three regions in Figure 1B constitutes a "black hole" attracting the triplet $\{G_1, G_2, G_3\}$ as a whole. Indeed, in that case, the conditional pairwise mutual information $I(G_1; G_2 \mid G_3)$ is larger than the actual mutual information $I(G_1; G_2)$, although the former appears as a subset of the latter in the Venn diagram of Figure 1B! If the entanglement $-I(G_1; G_2; G_3)$ is positive and significantly large, then this is an indication that one of these three genes may be cooperatively regulated by the other two, at least indirectly.

Figure 2 shows two illustrating examples. When the three-way mutual information is very high, then the three-dimensional scatter plot of the expression levels for the three genes tends to be restricted to a line, because the expression level of each of the genes tends to be, by itself, sufficient to predict the values of the other two, as, for example, is the case in coregulated genes. Indeed, the three genes for Figure 2A, *cheB*, *cheR*, and *cheY*, belong to the same *E. coli* chemotaxis operon, and their three-way mutual information was found very large and equal to $+1.24$ using the estimation technique described in Methods. Identifying high three-way mutual information triplets of genes may prove more powerful than using pairwise

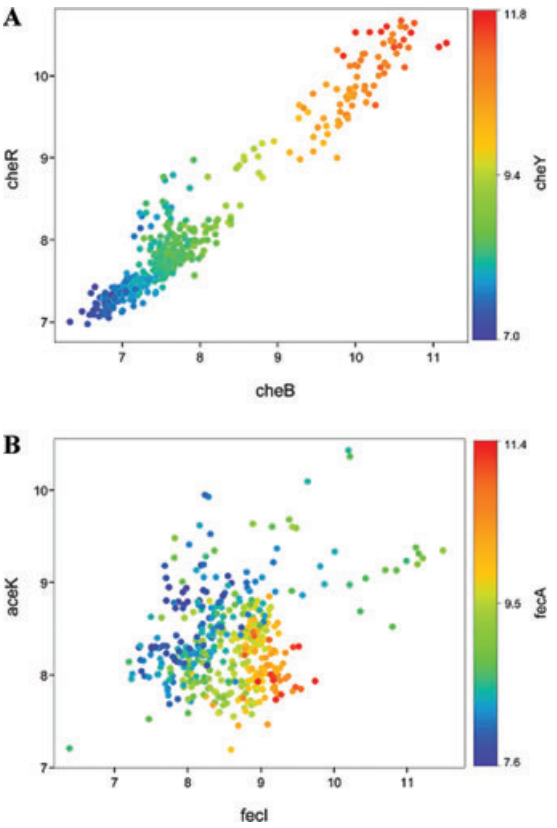

FIGURE 2. Examples of scatter plots illustrating high and low three-way mutual information. Expression data were downloaded from Ref. 12. (**A**) The expression of *cheY* as a function of the expression levels of *cheB* and *cheR*. The three genes belong to the same chemotaxis operon and are coregulated. (**B**) The expression of *fecA* as a function of the expression levels of *fecI* and *aceK*. The three genes are synergistically entangled with negative three-way mutual information, and the expression level of *fecA* is high only when simultaneously the expression level of *fecI* is high and the expression level of *aceK* is low.

correlations for clustering or biclustering genes into co-expressed modules.

On the other hand, Figure 2B shows an example of an "entangled" triplet of *E. coli* genes, *fecI, aceK,* and *fecA,* with negative three-way mutual information, equal to −0.16. It is seen that in some regions the expression level of *fecA* is not significantly associated with any one of the expression levels of the two other genes alone, but it is strongly associated with the two expression levels jointly. The corresponding "Boolean logic" can be described by the statement that high expression of *fecA* tends to occur only when simultaneously *fecI* is overexpressed and *aceK* is underexpressed, but not otherwise. This observation connecting these three particular genes is revisited in more detail in the discussion section below.

In previous work,[15,17] synergy values involving three genes were numerically estimated only for bilevel gene expression data, in other words, assuming genes are either "on" or "off," using arbitrary thresholds to binarize expression values inferred from microarrays. Here, as described in Methods, we directly evaluate them from the continuous expression levels.

Results and Discussion

Use of Three-Way Mutual Information for Gene Regulatory Network Inference

If the three-way mutual information among three genes is negative and its magnitude is significantly high, then this is an indication that one of the genes may be cooperatively regulated by the other two, at least partly or indirectly. In that case, it is reasonable to assume that the gene that is being regulated is the one that is most highly correlated with the pair of the other two genes, and in that case it can be easily proved that the pair of these two "regulating" genes have the lowest pairwise mutual information compared with the other two pairs. This observation suggests that the following quantity,

$$S(i,j) = \max_{\substack{k \\ \text{where } k \neq i,\ k \neq j,\\ I(G_i;G_k) < I(G_i;G_j)\\ I(G_i;G_k) < I(G_k;G_j)}} \left[-I(G_i;G_j;G_k)\right]$$

to which we refer as the "synergistic regulation index" can be used as a measure of the degree of confidence that gene G_i cooperatively regulates gene G_j. In other words, if we can identify a third gene serving as "synergistic partner" to gene G_i towards synergistically regulating gene G_j, then this will indicate that gene G_i is one of the regulators of gene G_j. In that case, it makes sense to assign a directed edge from node i to node j in the corresponding gene regulatory network.

There are many possible biological explanations for two genes G_i and G_k being members of a cooperative regulatory mechanism for gene G_j. In many cases, the relationship can be approximated by a Boolean logic function connecting the two regulating genes with the regulated gene. For example, such logical "AND gates" can be formed if G_i and G_k serve as two transcription factors with different binding sites on the promoter of G_j, if G_i and G_k form a dimer serving as transcription factor of G_j, or if G_i is a kinase required for the activation of G_k serving as transcription factor of G_j.

To provide an instructive example illustrating the capabilities of the synergistic regulation index to infer directed cooperative regulatory interactions, we synthesized gene expression values for 300 hypothetical microarray experiments involving a simple synthetic network consisting of 10 genes cooperatively regulated by AND and OR logic. The details of the synthetic network are described in Methods. There was no assumed prior knowledge about which genes play regulatory roles, that is, all 10 genes could be potential regulators of the other genes. The synergistic regulation index was significantly higher in all "regulator/regulated" ordered pairs of genes compared with all other ordered pairs of genes, thus successfully

identifying the full directed regulatory network, while pairwise mutual information–based techniques, including the CLR algorithm, were unable to correctly distinguish pairs of coregulated genes from real regulatory interactions in this example.

Given the inability of the synergistic regulation index to detect noncooperative interactions, or interactions of which the "synergistic partner" gene is missing from the expression dataset, we have selected a known pairwise mutual information–based methodology, the CLR algorithm,[12] to serve as a tool of detecting pairwise interaction, which we augment with our own complementary methodology, to arrive at an algorithm that we call the "synergy-augmented CLR algorithm."

Adaptive Background Correction

For each pair of genes G_i and G_j, where G_i is among the potential regulatory genes and G_j is among the potential target genes, the CLR algorithm evaluates the mutual information $M(i,j)$. The "background distribution" for this pair is constructed from two sets of mutual-information values: those corresponding to the ith row of the M matrix and those corresponding to the jth row of the M matrix. In other words the value of $M(i,j)$ is compared against the values of the mutual information between G_i and each of its potential target genes, as well as against the values of the mutual information between G_j and each of its potential regulatory genes. The two corresponding z-scores, z_i and z_j, for the two distributions are evaluated, and, if they are both non-negative, the final score for the interaction between G_i and G_j is equal to $\sqrt{z_i^2 + z_j^2}$, otherwise the score is equal to 0.

In our synergy-augmented CLR (SA-CLR) algorithm, we substitute $M(i, j)$ with the sum $M(i,j) + S(i,j)$ and we then follow precisely the same CLR background correction procedure on this sum, as described above. The quantity $M(i, j) + S(i, j)$ is asymmetric with respect to i and j even before undergoing background correction. The two components $M(i, j)$ and $S(i, j)$ serve complementary roles as evidenced by the equation $I(G_i; G_j) + [-I(G_i; G_j; G_k)] = I(G_i; G_j | G_k)$. Again, it is instructive to note that, when synergy is positive, this conditional mutual information is higher than the pairwise mutual information $I(G_i; G_j)$ despite the fact that the former appears as a subset of the latter in Figure 1B. The synergistic interaction index S provides complementary information to that of M, in the sense that the two quantities detect different aspects of true interactions.

Application to *E. coli* Gene Expression Dataset

We applied our methodology to a publicly available compendium of *E. coli* gene expression profiles,[12] combined with "ground truth" data consisting of 3,216 "transcription factor–target gene" regulatory interaction pairs, known from the RegulonDB database,[18] involving 153 transcription factors and 1,156 target genes, where the set of the 1,156 target genes included 100 of the 153 transcription factors. We used an expression matrix of Robust Multichip Average (RMA)-normalized expression values from 445 experiments using Affymetrix arrays with 7,231 probe sets. Both expression data and the RegulonDB data (version 4) were downloaded from Ref. 12.

Using the mutual information estimation technique described in Methods, we first evaluated a 153 × 1,156 matrix containing the values of the mutual information $M(i, j)$ between each potential transcription factor and each potential target gene. We also evaluated a 153 × 1,156 × 7,231 array containing the values of the three-way mutual information connecting each potential transcription factor, each potential target gene, and each potential synergistic partner out of a total of 7,231 microarray probes. We used the definition of $S(i, j)$ to extract the 153 × 1,156 matrix containing the corresponding synergistic regulation indexes. We then applied the adaptive background correction procedure for (A) $M(i, j)$ and for (B) $M(i, j) + S(i, j)$. In each case, we ranked the

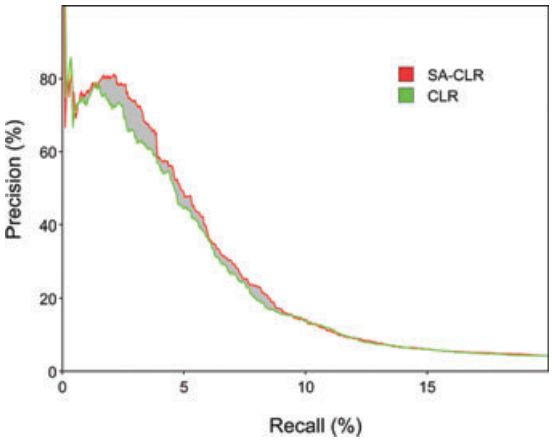

FIGURE 3. Comparison of precision-versus-recall curves after applying the CLR algorithm with synergy augmentation (SA-CLR, red line) and without (CLR, green line). The CLR algorithm was implemented using the data described in the text ("Application to *E. coli* gene expression dataset") and using the mutual information estimation method described in Methods.

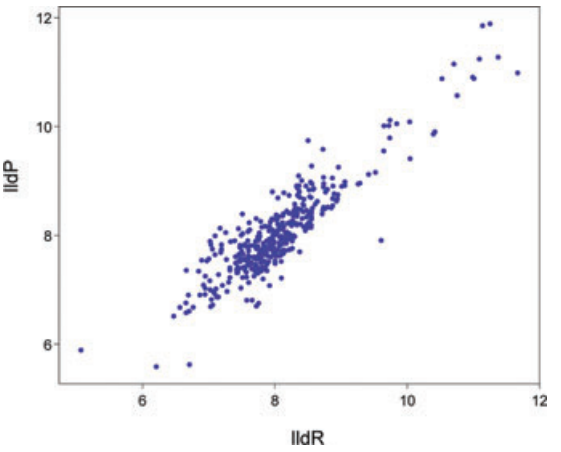

FIGURE 4. Example scatter plot of a typical inhibitory interaction. The transcription factor lldR is a known inhibitor of its own operon, including the gene lldP. The scatter plot misleadingly appears to indicate an excitatory interaction.

$153 \times 1{,}156 = 176{,}868$ gene pairs in terms of their score and we compared this ranking against the ground truth of the 3,216 known transcriptional interactions, therefore the probability of picking a correct interaction by pure chance would be 1.82%.

Figure 3 shows the corresponding two precision-versus-recall curves, labeled "CLR" and "SA-CLR" respectively, where precision (selectivity) is the fraction $TP/(TP + FP)$, recall (sensitivity) is the fraction $TP/(TP + FN)$ and TP, FP, and FN are the numbers of true positives, false positives, and false negatives, respectively, computed over a range of pruning thresholds. The improvement of SA-CLR over CLR is best demonstrated by the fact that among the top 150 SA-CLR predicted interactions, 106 where among the 3,216 "ground truth" interactions (precision = 70.7%). On the other hand, among the top 150 CLR predicted interactions, 96 were included in the "ground truth" interactions (precision = 64.0%). The areas under the precision-versus-recall curve for SA-CLR and CLR were 300.71 and 297.07, respectively. There was no need to provide comparisons against other methods such as regression and Bayesian networks, since this was already done in the CLR paper, demonstrating that CLR is highest-performing among those methods when using these data.

Performance in DREAM2 Challenge 5

The SA-CLR algorithm competed in the genome-scale network prediction challenge of the 2nd conference on Dialogue for Reverse Engineering Assessments and Methods (DREAM2), in which expression data for 3,456 genes under 300 different experimental conditions were prepared from both publicly available and private data. Both the conditions and the genes were disguised. Of the 3,456 genes, 320 were identified as transcription factors. The goal was to predict the regulatory targets of these transcription factors. The SA-CLR algorithm was the best performer in this challenge, with 67.5% precision at the point of the 100th correct prediction. The results are detailed in Table 1, in which GISL (Genomic Information Systems Laboratory) indicates our submission, while the other submissions are indicated as "Team 2" . . . "Team 5." For each submission, the DREAM2 committee computed the P values for the areas under the

TABLE 1. Results of the DREAM2 Genome-Scale Network Challenge

Team	1st	2nd	5th	20th	100th	500th	AU P/R Curve	AU ROC Curve	Combined $-\log_{10}(P \text{ value})$
GISL	1.000	1.000	0.714	0.690	0.676	0.036	0.059	0.611	40.5
Team 2	1.000	0.400	0.556	0.667	0.380	0.015	0.032	0.575	25.2
Team 3	1.000	1.000	0.625	0.769	0.515	0.024	0.047	0.572	24.1
Team 4	1.000	1.000	1.000	0.870	0.124	0.014	0.031	0.557	18.7
Team 5	0.500	0.500	0.333	0.087	0.024	0.009	0.010	0.527	10.0

The prediction accuracy after the 1st, 2nd, 5th, 20th, 100th, and 500th correct prediction is indicated, as well as the areas under the precision/recall curve and the receiver operating characteristic curve, indicated as AU P/R and AU ROC, respectively, as well as the final score for each submission (see text).

precision-versus-recall curve and under the receiver operating characteristic (ROC) curve, out of which a combined *P* value was computed, and the final score was evaluated as $-\log_{10}(P \text{ value})$.

Our submission was "unsigned," that is, without labeling regulatory interactions as excitatory or inhibitory. Predicting the type of interactions cannot lead to accurate results from merely the data given in this particular challenge without any indication of operon membership, time series, or specific conditions in the expression data. An example illustrating this point is presented in the scatter plot of Figure 4, in which the regulatory gene *lldR* is a repressor of its own operon, which also contains gene *lldP*. The shape of the scatter plot, while consistent with both genes being in the same operon, would misleadingly indicate that the interaction is excitatory.

The DREAM2 committee agreed to do an "extra-official" experiment, evaluating the score of our submission assuming, first, that *all* predicted interactions are excitatory and, second, that *all* predicted interactions are inhibitory. Remarkably, the combined $-\log_{10}(P \text{ value})$ scores were 26.4 and 26.5, respectively—higher than those of the other unsigned submissions, indicating the exceptional accuracy of the SA-CLR algorithm.

Discussion

We believe that systems biology is approaching the point of the paradigm shift at which analysis of biological data will result in significant discoveries of novel biological mechanisms. We hope that the concept of synergy will provide a valuable tool for multivariate analysis of biological data, helping toward this paradigm shift. Indeed, in addition to the improvement in pairwise regulatory interaction prediction accuracy, an important advantage of the synergy-augmented CLR algorithm is the greatly enhanced potential of identifying novel combinatorial interactions providing valuable clues for biological discoveries. Each identified interaction is accompanied by a most likely "partner gene." If the amount of measured synergy is significant, this provides an indication that the three genes may be members of a shared pathway, at least indirectly. The low *P* value observed in our second validation experiment (see Methods) indicates that some of these interactions have true biological significance. Although this may not necessarily be true for several gene triplets identified as "entangled," the biological clues resulting from some of the identified triplets, when coupled with additional biological knowledge, will lead to deciphering of more biological mechanisms compared to those that can be revealed by the inference of pairwise transcriptional regulatory networks alone. Synergy-based analysis is also applicable to mixed biological data including phosphoproteomics and SNP mutations, thus opening new dimensions for potential discovery. Here we mention one example stemming from applying our methodology to the *E. coli* data, when focusing on a particular interaction.

It is known that FecI is a sigma factor that, in its activated form, directs the RNA polymerase core enzyme to the promoter of the *fecA* ferric citrate transporter operon. In addition to this known FecI-*fecA* interaction, the CLR algorithm had identified[12] a novel PdhR-*fecA* interaction, where *pdhR* is a pyruvate-sensing regulatory gene. The novel interaction was confirmed with real-time quantitative PCR, determining that the *fecA* operon reached its highest level of induction only when the two chemicals citrate (known to increase the expression of *fecI*) and pyruvate (known to increase the expression of *pdhR*) were both present in high concentrations.

Triggered by this finding, we used the synergy-augmented CLR algorithm to search for the most synergistic partner of the FecI-*fecA* interaction, which turns out to be the gene *aceK*, the bifunctional isocitrate gehydrogenase kinase/phosphatase. Interestingly, as was the case with Pdhr, pyruvate binds directly with AceK.[19] The synergy of *fecI* and *aceK* with respect to *fecA* results from the AND-like logic (Fig. 2B) that high *fecA* expression occurs only in the simultaneous high expression of *fecI* and low expression of *aceK*. It is known that phosphorylation inactivates the enzyme isocitrate dehydrogenase (IDH), and dephosphorylation activates IDH. It is also known that inactivity of IDH (as observed in mutants of the IDH gene *icd*) inhibits ferric citrate transport even in the high presence of citrate,[20] that is, even at high expression levels of *fecI*. Therefore, we hypothesize that if the expression level of *aceK* is high, it acts as a kinase so that IDH is phoshorylated and deactivated, thus inhibiting ferric citrate transport.

On the other hand, it is also known that pyruvate inhibits the kinase activity and instead causes AceK to act as IDH phosphatase,[19,21] thus dephosphorylating (activating) IDH and allowing ferric citrate transport in the presence of citrate. When pyruvate is exhausted, IDH is again deactivated,[21] and ferric citrate transport is inhibited as a result. These facts are consistent with the previous finding that citrate and pyruvate are jointly associated with *fecA* induction, further confirming the coupling between metabolism and ferric citrate transport discovered by the use of the CLR algorithm, and suggesting that *aceK* is a gene directly involved in this coupling.

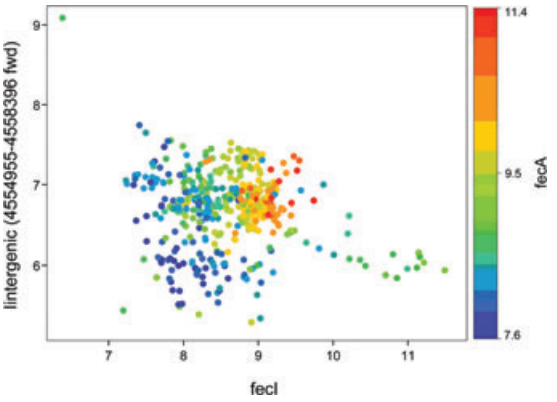

FIGURE 5. The expression of the unknown potential RNA gene corresponding to intergenic probe set IG_2826_4554955_4558396_fwd_at appears to be synergistically related to the expression of *fecI* with respect to the expression of *fecA*.

When searching over intergenic probes, we also identified another potential synergistic partner to the FecI-*fecA* interaction whose high expression, jointly with that of *fecI*, appears to be required for the high expression of *fecA* (Fig. 5). We speculate that the probe, IG_2826_4554955_4558396_fwd_at from the *E. coli* K12 complete genome NCBI accession NC_000913.2, contains a noncoding regulatory RNA gene involved in a shared related pathway.

Methods

Evaluation of Pairwise and Three-Way Mutual Information

For the comparison between SA-CLR and CLR presented in this paper, we used a three-dimensional extension of a mutual information estimator[22] that adaptively partitions the observation space based on the unknown underlying distributions of the samples. This method is

computationally efficient and therefore appropriate for performing the multiple permutation experiments required for the statistical analysis and validation of our experiments. For the DREAM 2 challenge, we used a three-dimensional extension of the same spline-based estimator[23] as was used with the CLR algorithm in Ref. 12, using seven bins in each dimension. This method divides observation space in to equally spaced bins and blurs the boundaries between the bins with spline basis functions using third-order B-splines. We found that the latter method yields slightly more accurate results at the expense of additional computational complexity.

Due to the need for estimating the three-way mutual information for all gene triplets, the SA-CLR algorithm requires the use of high-performance computing facilities. For example, in our case the estimation of the matrix with the values of the three-way mutual information took approximately one day of computation using 50 nodes of a computer cluster. Incorporating biological knowledge, combined with other techniques of dimensionality reduction such as biclustering[24,25] can be helpful toward both reducing computational requirement and increasing statistical significance.

Statistical Analysis and Validation of Experiments

The accuracy of algorithmic results, in our case, can be directly measured against a known ground truth, thus simplifying the problem of estimating their statistical significance. Because the aim of this paper is to establish the high performance of the regulatory interaction inference prediction performance of the SA-CLR algorithm, and in view of the fact that our method requires evaluating the three-way mutual information over a large number of gene triplets, we addressed the question of whether the observed improvement in transcriptional interaction prediction performance of the SA-CLR by adding the synergy component could be due to pure chance. If the improvement is not due to chance, then this suggests that SA-CLR genuinely captures the effect of additional synergistic biological mechanisms that are not captured by plain CLR. The only difference between the two algorithms is that SA-CLR uses the quantity $M(i, j) + S(i, j)$, while CLR uses just $M(i, j)$.

We performed two validation experiments: First, a "permutation experiment" in which we ran 1,000 permutations, in each of which the $153 \times 1{,}156$ values of the matrix S were randomly shuffled. For each such permuted matrix $\hat{S}$, we computed the background-normalized z-scores of $M + \hat{S}$, yielding a ranking of the gene pairs. We define the overall score as the area under the corresponding precision-versus-recall curve. The score of the SA-CLR algorithm was 300.7. The histogram of these scores is shown in Figure 6. The mean value extracted from the histogram was 292.8 and the standard deviation was 1.47, resulting in SA-CLR z-score of 5.4.

A potential argument against the above procedure is that a permuted S matrix may be mathematically incompatible with the fixed and related M matrix. Therefore, we performed a second "randomization experiment," as follows. The SA-CLR algorithm selects the "most synergistic" partner gene to generate the S matrix. If this partner gene is assigned randomly, rather than selecting the one that maximizes synergy, then this random choice will yield a new matrix $\hat{S}$. As before, $M + \hat{S}$ is background corrected yielding a new ranking of gene pairs and a corresponding overall score defined as the area underneath the precision-versus-recall curve. We created a new set of 1,000 such score values, yielding the corresponding histogram shown in Figure 6. The mean value extracted from that histogram was 290.0 and the standard deviation was 1.95, resulting in SA-CLR z-score of 5.5.

The advantage of the latter randomization methodology is that it establishes the significance of using the most synergistic, rather than a random, partner, suggesting a genuine biological significance for this choice.

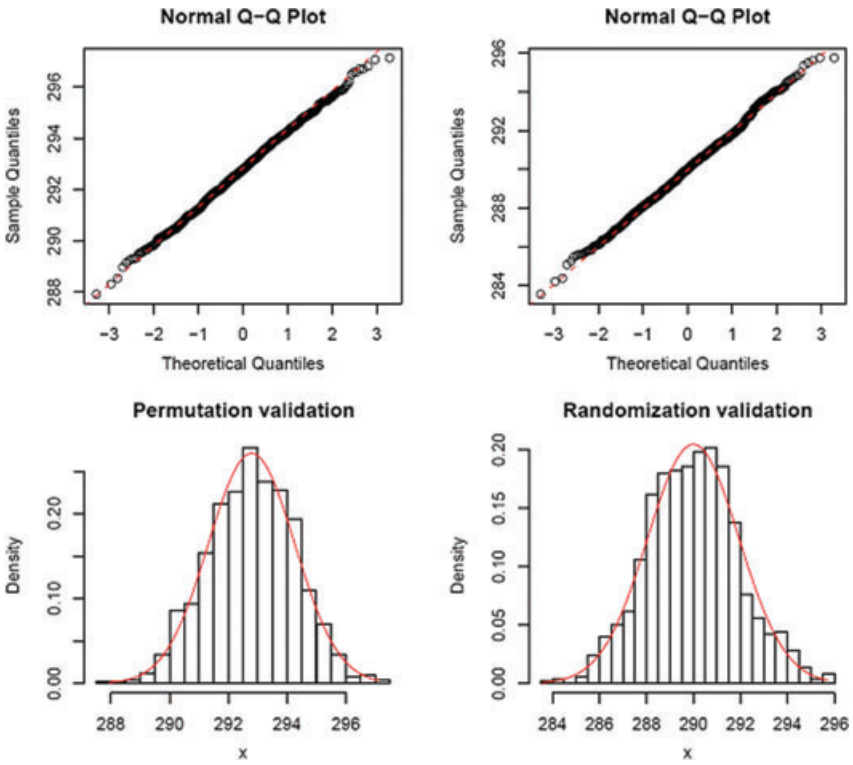

FIGURE 6. The histograms corresponding to the scores for the two validation experiments satisfy the Gaussian fit tests. The score for SA-CLR is 300.7, resulting in a z-score of 5.4 ($P = 3 \times 10^{-8}$) for the permutation validation and z-score of 5.5 ($P = 2 \times 10^{-8}$) for the randomization validation.

To examine the validity of using the normal distribution as a reference in determining P values for the histograms of Figure 6, we performed Kolmogorov-Smirnov tests against normal distributions with the same mean and variance. The P values for the permutation and randomization validation are 0.9275 and 0.7908, respectively, indicating no evidence to suggest that these two distributions are not normal. Therefore, the null hypothesis that they are from a normal distribution is accepted. As additional confirmation, we plotted the normal quantile-quantile (Q-Q) plots, containing the quantiles from the data set against those of a normal distribution. As seen in the figure, scores from both validation strategies form a distribution very close to normal, as points tend to form a straight line. We also show the histogram of the validation scores and the best normal fit, which also shows that the shape of the distributions is approximately that of a normal distribution.

With normality established, the z-scores mentioned above (5.4 and 5.5 for the permutation and randomization experiment, respectively) can be used to determine the P values, measuring the probability that the observed improvement of SA-CLR was due to chance. The P values for the permutation and randomization validation are then found to be 3×10^{-8} and 2×10^{-8}, respectively, thus validating the improvement.

Synthetic Network

We synthesized 300 hypothetical gene-expression experiments representing a simple gene regulatory network involving 10 genes $G_1, G_2, \ldots, G_{10}$, as in Figure 7. We assumed that genes G_5, G_6, G_7 are cooperatively coregulated by genes G_1, G_2 following AND logic,

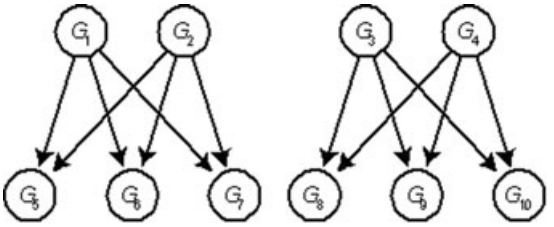

FIGURE 7. The synthetic regulatory network used to test the SA-CLR algorithm. SA-CLR correctly predicted this network, while the traditional relevance network approach incorrectly predicted interactions between coregulated genes.

and that genes G_8, G_9, G_{10} are cooperatively coregulated by genes G_3, G_4 following OR logic. We constructed the 10×300 gene-expression matrix using the NetBuilder facility.[26] The expression values of the four input genes were initially drawn from a uniform distribution in the interval (0, 1). To the resulting expression values of the six output genes we added random Gaussian noise with amplitude 0.1 and standard deviation 1. We then multiplied by 10 and added the value of 2, so that the values resemble typical normalized gene expression values.

We then analyzed the 10×300 expression matrix assuming no prior knowledge about the regulatory nature of any gene. The synergistic regulation indexes $S(i, j)$ fully recovered the directed network, identifying the six interactions scoring > 0.2.

On the other hand, the values of the pairwise mutual information cannot make this identification. The highest pairwise mutual information values occur between *coregulated* genes. The correct interactions have lower mutual information values.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Bansal, M. *et al*. 2007. How to infer gene networks from expression profiles. *Mol. Syst. Biol.* **3:** 78.
2. Margolin, A.A. *et al*. 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7**(Suppl 1): S7.
3. Butte, A.J. & I.S. Kohane. 2000. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing* 418–429.
4. Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers. San Francisco, CA.
5. Friedman, N. 2004. Inferring cellular networks using probabilistic graphical models. *Science* (New York, NY) **303:** 799–805.
6. Gardner, T.S. *et al*. 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* (New York, NY) **301:** 102–105.
7. Kishino, H. & P.J. Waddell. 2000. Correspondence analysis of genes and tissue types and finding genetic links from microarray data. *Genome Inform.* **11:** 83–95.
8. Schafer, J. & K. Strimmer. 2005. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* **21:** 754–764.
9. Cover, T.M. & J.A. Thomas. 2006. *Elements of information theory*. Wiley-Interscience. Hoboken, NJ.
10. Basso, K. *et al*. 2005. Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* **37:** 382–390.
11. Margolin, A.A. *et al*. 2006. Reverse engineering cellular networks. *Nat. Protocols* **1:** 662–671.
12. Faith, J.J. *et al*. 2007. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **5:** e8.
13. Schneidman, E. *et al*. 2006. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440:** 1007–1012.
14. Wang, K. *et al*. 2006. Genome-wide discovery of modulators of transcriptional interactions in human B lymphocytes. In *RECOMB 2006, LNBI 3909*. A. Apostolico, *et al*., Eds. Springer, Berlin/Heidelberg, pp. 348–362.
15. Anastassiou, D. 2007. Computational analysis of the synergy among multiple interacting genes. *Mol. Syst. Biol.* **3:** 83.
16. Watkinson, J. *et al*. 2008. Identification of gene interactions associated with disease from gene expression data using synergy networks. *BMC Syst. Biol.* **2:** 10. [Highly accessed].
17. Varadan, V., D.M. Miller, 3rd & D. Anastassiou. 2006. Computational inference of the molecular logic for synaptic connectivity in C. elegans. *Bioinformatics* **22:** e497–e506.
18. Salgado, H. *et al*. 2006. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory

network, operon organization, and growth conditions. *Nucleic Acids Res.* **34:** D394–D397.
19. Miller, S.P. *et al*. 2000. Locations of the regulatory sites for isocitrate dehydrogenase kinase/phosphatase. *J. Biol. Chem.* **275:** 833–839.
20. Braun, V. 1997. Surface signaling: novel transcription initiation mechanism starting from the cell surface. *Arch. Microbiol.* **167:** 325–331.
21. el-Mansi, E.M., H.G. Nimmo & W.H. Holms. 1986. Pyruvate metabolism and the phosphorylation state of isocitrate dehydrogenase in Escherichia coli. *J. Gen. Microbiol.* **132:** 797–806.
22. Darbellay, G.A. & I. Vajda. 1999. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Trans. Inform. Theory* **45:** 1315–1321.
23. Daub, C.O. *et al*. 2004. Estimating mutual information using B-spline functions–an improved similarity measure for analysing gene expression data. *BMC Bioinform.* **5:** 118.
24. Bonneau, R. *et al*. 2006. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.* **7:** R36.
25. Hayete, B., T.S. Gardner & J.J. Collins. 2007. Size matters: network inference tackles the genome scale. *Mol. Syst. Biol.* **3:** 77.
26. Wegner, K. *et al*. 2007. The 'NetBuilder' project: development of a tool for constructing, simulating, evolving, and analysing complex regulatory networks. *BMC Syst. Biol.* **1**(Suppl 1): P72.